Introduction to Sampling and Data

class="introduction" We encounter statistics in our daily lives more often than we probably realize and from many different sources, like the news. (credit: David Sim)

Chapter objective

By the end of this chapter, the student should be able to:

- Recognize and differentiate between key terms.
- Apply various types of sampling methods to data collection.

You are probably asking yourself the question, "When and where will I use statistics?" If you read any newspaper, watch television, or use the Internet, you will see statistical information. There are statistics about crime, sports, education, politics, and real estate. Typically, when you read a newspaper article or watch a television news program, you are given sample information. With this information, you may make a decision about the correctness of a statement, claim, or "fact." Statistical methods can help you make the "best educated guess."

Since you will undoubtedly be given statistical information at some point in your life, you need to know some techniques for analyzing the information thoughtfully. Think about buying a house or managing a budget. Think about your chosen profession. The fields of economics, business, psychology, education, biology, law, computer science, police science, and early childhood development require at least one course in statistics.

Included in this chapter are the basic ideas and words of probability and statistics. You will soon understand that statistics and probability work together. You will also learn how data are gathered and what "good" data can be distinguished from "bad."

Some Key Statistical Terms and Definitions

The science of **statistics** deals with the collection, analysis, interpretation, and presentation of **data**.

The process of statistical analysis follows these broad steps.

1. Defining the problem
2. Planning the study
3. Collecting the data for the study
4. Analysis of the data
5. Interpretations and conclusions based on the analysis

For example, we may wonder if there is a gap between how much men and women are paid for doing the same job. This would be the problem we want to investigate. Before we do the investigation, we would want to spend some time defining the problem. This could include defining terms (e.g. what do we mean by “paid”? what constitutes the “same job”?). Then we would want to state a **research question**. A research question is the overarching question that the study aims to address. In this example, our research question might be: “Does the gender wage gap exist?”.

Once we have the problem clearly defined, we need to figure out how we are going to study the problem. This would include determining how we are going to collect the data for the study. Since it is unlikely we are going to find out the salary and position of every employee in the world (i.e. the population), we need to instead collect data from a subset of the whole (i.e. a sample). The process of how we will collect the data is called the **sampling technique**. The overall plan of how the study is designed is called the **sampling design** or **methodology**.

Once we have the methodology, we want to implement it and collect the actual data.

When we have the data, we will learn how to organize and summarize data. Organizing and summarizing data is called **descriptive statistics**. Two ways to summarize data are by visually summarizing the data (for example, a histogram) and by numerically summarizing the data (for example, the average). After we have summarized the data, we will use formal methods for drawing conclusions from "good" data. The formal methods are called **inferential statistics**. Statistical inference uses probability to determine how confident we can be that our conclusions are correct.

Once we have summarized and analyzed the data, we want to see what kind of conclusions we can draw. This would include attempting to answer the research question and recognizing the limitations of the conclusions.

In this course, most of our time will be spent in the last two steps of the statistical analysis process (i.e. organizing, summarizing and analyzing data). To understand the process of making inferences from the data, we must also learn about probability. This will help us understand the likelihood of random events occurring.

Key Idea and Terms

In statistics, we generally want to study a **population**. You can think of a population as a collection of persons, things, or objects under study. You can think of a population as a collection of persons, things, or objects under study. The person, thing or object under study (i.e. the object of study) is called the **observational unit**. What we are measuring or observing about the observational unit is called the **variable**. We often use the letters X or Y to represent a variable. A specific instance of a variable is called **data**.

Suppose our research question is “Do current NHL forwards who make over $3 million a year score, on average, more than 20 points a season?”
The population would be all of the NHL forwards who make over $3 million a year and who are currently playing in the NHL. The observational

unit is a single member of the population, which would be any forward that made over $3 million year. The variable is what we are studying about the observation unit, which is the number points a forward in the population gets in a season. A data value would be the actual number of points.

In the above example, it would be reasonable to look at the population when doing the statistical analysis as the population is very well defined, there are many websites that have this information readily available, and the population size is relatively small. But this is not always the case. For example, suppose you want to study the average profits of oil and gas companies in the world. This might be very hard to get a list of all of the oil and gas companies in the world and get access to their financial reports. When the population is not easily accessible, we instead look at a **sample**. The idea of **sampling** (the process of collecting the sample) is to select a portion (or subset) of the larger population and study that portion (the sample) to gain information about the population.

Because it takes a lot of time and money to examine an entire population, sampling is a very practical technique. If you wished to compute the overall grade point average at your school, it would make sense to select a sample of students who attend the school. The data collected from the sample would be the students' grade point averages. In federal elections, opinion poll samples of 1,000–2,000 people are taken. The opinion poll is supposed to represent the views of the people in the entire country. Manufacturers of canned carbonated drinks take samples to determine if a 16 ounce can contains 16 ounces of carbonated drink.

It is important to note that though we might not know the population, when we decide to sample from it, it is fairly static. Going back to the example of the NHL forwards, if we were to gather the data for the population right now that would be our fixed population. But if you took a sample from that population and your friend took a sample from that population, it is not surprising that you and your friend would get a different sample. That is, there is one population, but there are many, many different samples that can be drawn from the sample. How the samples vary from each other is called **sampling variability**. The idea of sampling variability is a key concept in statistics and we will come back to it over and over again.

Data is plural. Datum is singular.

As mentioned above, a variable, or random variable, notated by capital letters such as X and Y, is a characteristic of interest for each person or thing in a population. Data are the actual values of the variable. Data and variables fall into two general types: either they are measuring something and they are not measuring. When a variable is measuring or counting something, it is called a **quantitative** variable and the data is called quantitative. When a variable is not measuring or counting something, it is called a **categorical** variable and the data is called categorical data. For a variable to be considered quantitative, the distance between each number has to be fixed. In general, quantitative variables measure something and take on values with equal units such as weight in pounds or number of people in a line. Categorical variables place the person or thing into a category such as colour of car or opinion on topic.

- In the NHL forwards example, the variable is quantitative as we investigating the number of points a player has.
- In the gender gap example, there were three variables: the salary, gender, and the position. The salary is a quantitative variable as we are investigating the amount people make. Gender is a categorical variable as we are categorizing someone's gender. Position is also categorical

as we are categorizing their type of employment.

- Sometimes though determining the type of a variable (i.e. quantitative or categorical) is not always cut and dry. In particular, **Likert scales** or rating scales are tricky to place. A Likert scale is any scale where you are asked to state your opinion on a scale. For example, you may be asked whether you strongly agree, agree, neutral, disagree or strongly disagree with a statement. Sometimes there is a number associated with the rating. For example, write 5 if you strongly agree and 1 if you strongly disagree. Technically, a Likert scale is a categorical data as we are categorizing people’s opinions and the number is just a short form for the category.

When you are asked to categorize the data or variable, first determine what the observation unit is. Then determine the variable being studied. Then think about what the data will look like. If the data is a number, then it is usually quantitative data (be wary of Likert scales). If the data is word or category, then it is categorical data.

For the following research questions, state the observational unit, the variable being studied, and the type of variable.

1. What is the average monthly temperature in Edmonton?
2. What is the highest belt colour that most students of karate earn in Canada?
3. What is the average weight of greyhound dogs?
4. What is the average gross profit of movies made in 2016?
5. What is the average user rating of Jessica Jones season 1 on IMDB?
6. What is the most common colour of car in Nova Scotia?

---

1. Observational unit: Edmonton. Variable: Monthly Temperature. Type: Quantitative.
2. Observational unit: Student of karate in Canada. Variable: Highest colour of belt earned. Type: Categorical.
3. Observational unit: Greyhounds. Variable: Weight. Type: Quantitative.
4. Observational unit: Movies made in 2016. Variable: Gross profit. Type: Quantitative.
5. Observational unit: Jessica Jones. Variable: User ratings. Type: Categorical.
6. Observational unit: Cars in Nova Scotia.

Variable: Colour. Type: Categorical.

Two words that come up often in statistics are **mean** and **proportion**. These are two example of numerical descriptive statistics. If you were to take three exams in your math classes and obtain scores of 86, 75, and 92, you would calculate your mean score by adding the three exam scores and dividing by three (your mean score would be 84.3 to one decimal place). If, in your math class, there are 40 students and 22 are men then the proportion of men in the course is 55% and the proportion of women is 45%.

From the sample data, we can calculate a statistic. A **statistic** is a numerical summary that represents a property of the sample. For example, if we consider one math class to be a sample of the population of all math classes, then the mean number of points earned by students in that one math class at the end of the term is an example of a statistic. The statistic is an estimate of a population parameter, in this case the mean. A **parameter** is a numerical summary that represents a property of the population. Since we considered all math classes to be the population, then the mean number of points earned per student over all the math classes is an example of a parameter (i.e. the population mean). If we took a sample of students from the math class and found the mean points earned per student in the sample, then we would have found a statistic (i.e. the sample mean).

In the NHL example, a sample of the population may be 31 forwards who make over $3 million per year. The sample was chosen by randomly choosing one forward who makes over $3 million from each team (if you are reading this after Sept. 2021, this would be changed to 32). The process of choosing the sample is called sampling. We would then collect the data for the sample, which would be the number of points each player in our sample gets in one season. The statistic would be the mean of the total number of points for the sample. The parameter at this point would be unknown, but we could estimate it with our statistic. To find the parameter, we would have to find the mean of the total number of points for the population.

One of the main concerns in the field of statistics is how accurately a statistic estimates a parameter. The accuracy really depends on how well the sample represents the population. The sample must contain the characteristics of the population in order to be a representative sample. We are interested in both the sample statistic and the population parameter in inferential statistics. In a later chapter, we will use the sample statistic to test the validity of the established population parameter.

Determine what the key terms refer to in the following study. We want to know what proportion of first-year students get to ABC college using public transit. We randomly survey 100 first year students at ABC college.

The **population** is all first year students attending ABC college this term.

The **sample** depends on how we choose the students. One possible answer could be all students enrolled in one section of a beginning statistics course at ABC College (although this sample would not be deemed random nor representative of the entire population).

The **variable** would be whether a first-year student uses public transportation to get to ABC college or not.

The **data** are the actual values of the variable. As students would either use public transportation or not, the data would be "yes" or "no, or "public transporation" or "not public transportation" (depending on how you chose to represent your data).

The **statistic** is the proportion of students in your SAMPLE who use public transportation to get to ABC college. (Note: The mean would not be an appropriate summary here as you cannot find the mean of categorical data).

The **parameter** is the proportion of ALL first-year students who use public transportation to get to ABC college.

Try It

Determine what the key terms refer to in the following study. We want to know the average (mean) amount of money spent on school uniforms each year by families with children at Knoll Academy. We randomly survey 100 families with children in the school. Three of the families spent $65, $75, and $95, respectively.

Try It Solutions

The **population** is all families with children attending Knoll Academy.

The **sample** is a random selection of 100 families with children attending Knoll Academy.

The **parameter** is the average (mean) amount of money spent on school uniforms by families with children at Knoll Academy.

The **statistic** is the average (mean) amount of money spent on school uniforms by families in the sample.

The **variable** is the amount of money spent by one family. Let X = the amount of money spent on school uniforms by one family with children attending Knoll Academy.

The **data** are the dollar amounts spent by the families. Examples of the data are $65, $75, and $95.

Determine what the key terms refer to in the following study.

A study was conducted at a local college to analyze the average cumulative GPA's of students who graduated last year. Fill in the letter of the phrase that best describes each of the items below.

1.___ Population 2.___ Statistic 3.___ Parameter 4.___ Sample 5.___ Variable 6.___ Data

- a) all students who attended the college last year
- b) the cumulative GPA of one student who graduated from the college last year
- c) 3.65, 2.80, 1.50, 3.90
- d) a group of students who graduated from the college last year, randomly selected
- e) the average cumulative GPA of students who graduated from the college last year
- f) all students who graduated from the college last year
- g) the average cumulative GPA of students in the study who graduated from the college last year

1. f 2. g 3. e 4. d 5. b 6. c

Determine what the key terms refer to in the following study.

As part of a study designed to test the safety of automobiles, the National Transportation Safety Board collected and reviewed data about the effects of an automobile crash on test dummies. Here is the criterion they used:

Speed at which Cars Crashed	Location of “drive” (i.e. dummies)
35 miles/hour	Front Seat

Cars with dummies in the front seats were crashed into a wall at a speed of 35 miles per hour. We want to know the proportion of dummies in the driver’s seat that would have had head injuries, if they had been actual drivers. We start with a simple random sample of 75 cars.

The **population** is all cars containing dummies in the front seat.

The **sample** is the 75 cars, selected by a simple random sample.

The **parameter** is the proportion of driver dummies (if they had been real people) who would have suffered head injuries in the population.

The **statistic** is proportion of driver dummies (if they had been real people) who would have suffered head injuries in the sample.

The **variable** $X =$ the number of driver dummies (if they had been real people) who would have suffered head injuries.

The **data** are either: yes, had head injury, or no, did not.

Determine what the key terms refer to in the following study.

An insurance company would like to determine the proportion of all medical doctors who have been involved in one or more malpractice lawsuits. The company selects 500 doctors at random from a professional directory and determines the number in the sample who have been involved in a malpractice lawsuit.

The **population** is all medical doctors listed in the professional directory.

The **parameter** is the proportion of medical doctors who have been involved in one or more malpractice suits in the population.

The **sample** is the 500 doctors selected at random from the professional directory.

The **statistic** is the proportion of medical doctors who have been involved in one or more malpractice suits in the sample.

The **variable** $X =$ the number of medical doctors who have been involved in one or more malpractice suits.

The **data** are either: yes, was involved in one or more malpractice lawsuits, or no, was not.

References

The Data and Story Library, http://lib.stat.cmu.edu/DASL/Stories/CrashTestDummies.html (accessed May 1, 2013).

Chapter Review

The mathematical theory of statistics is easier to learn when you know the language. This module presents important terms that will be used throughout the text.

HOMEWORK

For each of the following eight exercises, identify: a. the population, b. the sample, c. the parameter, d. the statistic, e. the variable, and f. the data. Give examples where appropriate.

> A fitness center is interested in the mean amount of time a client exercises in the center each week.

> Ski resorts are interested in the mean age that children take their first ski and snowboard lessons. They need this information to plan their ski classes optimally.

---

1. all children who take ski or snowboard lessons

2. a group of these children
3. the population mean age of children who take their first snowboard lesson
4. the sample mean age of children who take their first snowboard lesson
5. X = the age of one child who takes his or her first ski or snowboard lesson
6. values for X, such as 3, 7, and so on

A cardiologist is interested in the mean recovery period of her patients who have had heart attacks.

Insurance companies are interested in the mean health costs each year of their clients, so that they can determine the costs of health insurance.

---

1. the clients of the insurance companies
2. a group of the clients
3. the mean health costs of the clients
4. the mean health costs of the sample
5. X = the health costs of one client
6. values for X, such as 34, 9, 82, and so on

A politician is interested in the proportion of voters in his district who think he is doing a good job.

A marriage counselor is interested in the proportion of clients she counsels who stay married.

---

1. all the clients of this counselor
2. a group of clients of this marriage counselor
3. the proportion of all her clients who stay married
4. the proportion of the sample of the counselor's clients who stay married
5. $X =$ the number of couples who stay married
6. yes, no

Political pollsters may be interested in the proportion of people who will vote for a particular cause.

A marketing company is interested in the proportion of people who will buy a particular product.

---

1. all people (maybe in a certain geographic area, such as the United States)
2. a group of the people
3. the proportion of all people who will buy the product
4. the proportion of the sample who will buy the product
5. $X =$ the number of people who will buy it
6. buy, not buy

Use the following information to answer the next three exercises: A Lake Tahoe Community College instructor is interested in the mean number of days Lake Tahoe Community College math students are absent from class during a quarter.

What is the population she is interested in?

1. all Lake Tahoe Community College students
2. all Lake Tahoe Community College English students
3. all Lake Tahoe Community College students in her classes
4. all Lake Tahoe Community College math students

Consider the following:

X = number of days a Lake Tahoe Community College math student is absent

In this case, *X* is an example of a:

1. variable.
2. population.
3. statistic.
4. data.

---

a

The instructor's sample produces a mean number of days absent of 3.5 days. This value is an example of a:

1. parameter.
2. data.
3. statistic.
4. variable.

Glossary

Average
: also called mean or arithmetic mean; a number that describes the central tendency of

the data

Categorical Variable

variables that take on values that are names or labels

Data

a set of observations (a set of possible outcomes); most data used in statistical research can be put into two groups: **categorical** (an attribute whose value is a label) or **quantitative** (an attribute whose value is indicated by a number). Categorical data can be separated into two subgroups: **nominal** and **ordinal**. Data is nominal if it cannot be meaningfully ordered. Data is ordinal if the data can be meaningfully ordered. Quantitative data can be separated into two subgroups: **discrete** and **continuous**. Data is discrete if it is the result of counting (such as the number of students of a given ethnic group in a class or the number of books on a shelf). Data is continuous if it is the result of measuring (such as distance traveled or weight of luggage)

Numerical Variable

variables that take on values that are indicated by numbers

Parameter

a number that is used to represent a population characteristic and that generally cannot be determined easily

Population
all individuals, objects, or measurements whose properties are being studied

Probability
a number between zero and one, inclusive, that gives the likelihood that a specific event will occur

Proportion
the number of successes divided by the total number in the sample

Representative Sample
a subset of the population that has the same characteristics as the population

Sample
a subset of the population studied

Statistic
a numerical characteristic of the sample; a statistic estimates the corresponding population parameter.

Variable
a characteristic of interest for each person or object in a population

Data, Sampling and Variation

Data may come from a population or from a sample. Small letters like x or y generally are used to represent data values. Most data can be put into the following categories:

- Categorical
- Quantitative

Categorical data (also called qualitative data) are the result of categorizing or describing attributes of a population. Hair colour, blood type, ethnic group, the car a person drives, and the street a person lives on are examples of categorical data. Categorical data are generally described by words or letters. For instance, hair colour might be black, dark brown, light brown, blonde, grey, or red. Blood type might be AB+, O-, or B+. Researchers often prefer to use quantitative data over categorical data because it lends itself more easily to mathematical analysis. For example, it does not make sense to find an average hair or colour or blood type.

There are two types of categorical data: nominal and ordinal. **Nominal data** is categorical data that cannot be ordered in a meaningful way. For example, the colour of a car is categorical, but the order of the colours are not meaningful. **Ordinal data** is categorical data that can be ordered in a meaningful way. For example, the level of satisfaction someone has with their experience at a restaurant from not at all satisfied to completely satisfied.

Quantitative data are always numbers. Quantitative data are the result of **counting** or **measuring** attributes of a population. Amount of money, pulse rate, weight, number of people living in your town, and number of students who take statistics are examples of quantitative data. Quantitative data may be either **discrete** or **continuous**.

All data that are the result of counting are called **quantitative discrete data**. These data take on only certain numerical values. If you count the number of phone calls you receive for each day of the week, you might get values such as zero, one, two, or three.

All data that are the result of measuring are **quantitative continuous data** assuming that we can measure accurately. Measuring time, distance, area, and so on; anything that can be subdivided and then subdivided again and again is a continuous variable. If you and your friends carry backpacks with books in them to school, the numbers of books in the backpacks are discrete data and the weights of the backpacks are continuous data.

Data Sample of Quantitative Discrete Data

The data are the number of books students carry in their backpacks. You sample five students. Two students carry three books, one student carries four books, one student carries two books, and one student carries one book. The numbers of books (three, four, two, and one) are the quantitative discrete data.

Try It

The data are the number of machines in a gym. You sample five gyms. One gym has 12 machines, one gym has 15 machines, one gym has ten machines, one gym has 22 machines, and the other gym has 20 machines. What type of data is this?

Try It Solutions

quantitative discrete data

Data Sample of Quantitative Continuous Data

The data are the weights of backpacks with books in them. You sample the same five students. The weights (in pounds) of their backpacks are 6.2, 7, 6.8, 9.1, 4.3. Notice that backpacks carrying three books can have different weights. Weights are quantitative continuous data because weights are measured.

Try It

The data are the areas of lawns in square feet. You sample five houses. The areas of the lawns are 144 sq. feet, 160 sq. feet, 190 sq. feet, 180 sq. feet, and 210 sq. feet. What type of data is this?

Try It Solutions

quantitative continuous data

You go to the supermarket and purchase three cans of soup (19 ounces) tomato bisque, 14.1 ounces lentil, and 19 ounces Italian wedding), two packages of nuts (walnuts and peanuts), four different kinds of vegetable (broccoli, cauliflower, spinach, and carrots), and two desserts (16 ounces Cherry Garcia ice cream and two pounds (32 ounces chocolate chip cookies).

Name data sets that are quantitative discrete, quantitative continuous, categorical ordinal, and categorical nominal.

One Possible Solution:

- The three cans of soup, two packages of nuts, four kinds of vegetables and two desserts are quantitative discrete data because you count them.
- The weights of the soups (19 ounces, 14.1 ounces, 19 ounces) are quantitative continuous data because you measure weights as precisely as possible.
- Types of soups, nuts, vegetables and desserts are categorical nominal data because they are categories and fundamentally words. Further, there is no meaningful order.
- Descriptions of amount of rain (e.g. light, heavy) are categorical ordinal data as they categories but have a meaningful order.

Try to identify additional data sets in this example.

The data are the colors of backpacks. Again, you sample the same five students. One student has a red backpack, two students have black backpacks, one student has a green backpack, and one student has a gray backpack. The colors red, black, black, green, and gray are categorical nominal data.

Try It

The data are the colors of houses. You sample five houses. The colors of the houses are white, yellow, white, red, and white. What type of data is this?

Try It Solutions

categorical nominal data

Note

You may collect data as numbers and report it categorically. For example, the quiz scores for each student are recorded throughout the term. At the end of the term, the quiz scores are reported as A, B, C, D, or F. The data is ordinal as there is a meaningful order.

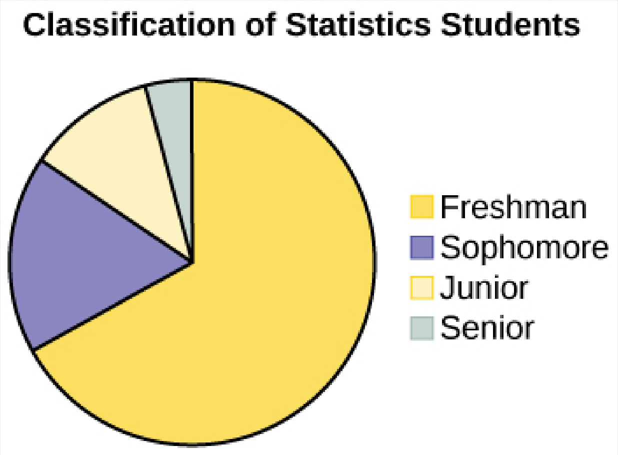
Try It

Determine the correct data type (quantitative or categorical) for the number of cars in a parking lot. Indicate whether quantitative data are continuous or discrete.

Try It Solutions

quantitative discrete

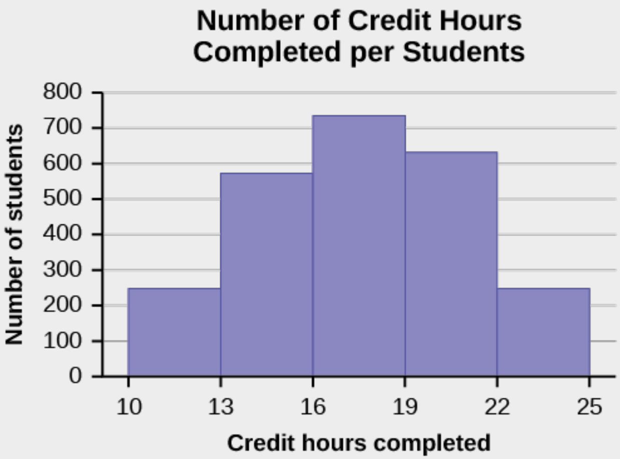
A statistics professor collects information about the classification of her students as freshmen, sophomores, juniors, or seniors. The data she collects are summarized in the pie chart [link]. What type of data does this graph show?



This pie chart shows the students in each year, which is **categorical nominal data**.

Try It

The registrar at State University keeps records of the number of credit hours students complete each semester. The data he collects are summarized in the histogram. The class boundaries are 10 to less than 13, 13 to less than 16, 16 to less than 19, 19 to less than 22, and 22 to less than 25.



What type of data does this graph show?

Try It Solutions

A histogram is used to display quantitative data: the numbers of credit hours completed. Because students can complete only a whole number of hours (no fractions of hours allowed), this data is quantitative discrete.

Sampling

Gathering information about an entire population often costs too much or is virtually impossible.

Instead, we use a sample of the population. The goal would be to use information from the sample to estimate information about the population.

To collect the sample, a sampling technique is used. Not all sampling techniques are created equal, though. A good sampling technique meets the following criteria:

- The sample is collected randomly
- The sample is representative of the population
- The size of the sample is large enough

If a sampling technique does not meet these criteria, then it is not appropriate to make inferences from the data. For example, it would not be appropriate to estimate the population mean from the sample mean.

A random sample reduces bias, promotes representativeness, and is a key component to sampling. **To do any scientific statistical analysis on sample data, the sample has to be randomly selected**. In a **random sample**, members of the population are selected in such a way that each has an equal chance of being selected. To ensure that a sample is collected randomly, some element of randomness needs to be included in the sampling technique. This can involve using dice to choose the time to start collecting data or using a random number generator to pick names from a list of names.

Humans in general are not very random. Therefore, the randomness added to the sampling technique cannot be someone “randomly” choosing something. The randomness has to come from a random event (like rolling dice, flipping a coin, using a random number generator).

A sample is **representative** if it shares similar characteristics to the population. For example, suppose that the students at a university are distributed as follows by faculty:

- Business: 20%
- Arts: 25%
- Science and Engineering: 30%
- Nursing: 15%
- Education: 10%

Then a sample would be representative of this population if the distribution of the students’ faculty in the sample was similar to the population. It doesn’t have to be exactly the same, but it should be close. A random sample will generate a fairly representative sample, but it doesn’t guarantee it.

What makes a sample representative depends on what is being studied. For example, if we are looking at the average age of students at a university, making sure we get students from each faculty would be important, but making sure we get students from various political affiliations might not be.

Determining if a sample is **large enough** is a bit arbitrary and depends on the situation. In general, the larger the sample size the better, but issues such as time and money need to be taken into account. You don't want to interview 5000 people, when 50 people would do. In Chapter 7, we will look at a formula that determines how many members of a population need to be in a sample depending on the level of error we are comfortable with. Until then, as a general rule, if the data is quantitative, a sample of at least 30 is usually good enough. While if the data is categorical, a sample of at least 100 is usually good enough.

In general, even if a sample is collected extremely well, it will not be perfectly representative of the population. The discrepancy between the sample and the population is called **chance error due to sampling**. When dealing with samples, there will always be error. Statistics helps us to understand and even measure this error. As a rule, the larger the random sample, in general the smaller the sampling error.

Generally, a sample that is collected randomly will likely be representative. But this is not guaranteed. For example, it is possible to collect a random sample of university students that happens to only contain students from one faculty. It is unlikely but possible.

A large sample size does not guarantee a representative sample. Nor does a small sample size guarantee a non-representative sample. To illustrate, a sample of ten university students could be chosen so that proportion of students from each gender in the sample is similar to the population, and the proportion of students from each faculty in the sample is similar to the population. Thus, the sample of size 10 would be representative. The point of a larger sample size is that the larger the sample, the more likely it is to be representative.

Of the three characteristics of a good sample, the most important one for statistical analysis is that the sample is collected randomly.

Areas of concern for sampling bias

When people publish their research, they include a description of their sampling technique. This is called the methodology. When evaluating a sampling technique, check to see if the sample was collected randomly, if it is representative of the population, and if the sample is large enough. Here are some examples of areas of concern when looking at methodologies:

1. *Undercoverage* occurs when a particular subset of the population is excluded from the process of selecting the sample. For example, if no one from the faculty of nursing is included in the sample, then we would say that the faculty of nursing is undercovered. As another example, undercoverage has been a specific concern in drug research over the years. In particular, women have been traditionally excluded from drug studies because of their menstrual cycles, but this results in the research only indicating how well the drug works for men.
2. *Nonresponse bias* occurs when a member of the population that is selected as part of the sample cannot be contacted or refuses to participate. Have you ever refused to be part of a telephone study? If so, you are contributing to nonresponse bias.

    - Similar to nonresponse bias is *voluntary response bias*. Here a large segment of the population is contacted and people choose to participate or not. Examples of this are mail-out surveys or online polls. In these

    situations, usually the person is very invested in the issue so that is why they take the time to answer. This results in non-representative samples. Another form of voluntary response bias is online surveys. Here, only people familiar with the website are likely to participate or "volunteer" to be part of the survey.
  - *Response rate* is a measure of how many people responded out of the total contacted. If the response rate is low, then this suggests a very narrow segment of the population answered. This would raise concerns about representativeness.

5. Asking potentially awkward questions might result in untruthful responses. This is called *response bias*. For example, if you are asked if you have ever had a sexually transmitted infection, you may not want to divulge that. One way to minimize response bias is to allow participants in a study to answer the questions anonymously.
6. Improper wording of questions being asked might result in skewed answers. Here is an example of a question that skews the results:

   - Do you think it should be easier for seniors to make ends meet?

     ○ Yes – they've worked hard and helped

build our country

○ No – seniors don’t need any help or recognition

> The wording of this question makes it hard to say "no". Thus, skewing the results towards "yes".

A famous example of a survey that had a very poor methodology was the incorrect prediction by the Literary Digest that Dewey would beat Truman in the 1936 US election. Check out the following website for more information: https://www.math.upenn.edu/~deturck/m170/wk4/lecture/case2.html

Sampling techniques

Most statisticians and researchers use various methods of random sampling in an attempt to achieve a good sample. This section will describe a few of the most common techniques: simple random sampling, (proportional) stratified random sampling, cluster sampling, systematic random sampling, and convenience sampling.

Simple random sampling

The easiest method to describe is called a **simple random sample**. In this technique, a random sample is taken from the members of the population. This can be done by putting the names (or identifier) of all members of the population into a hat and pulling out those names (or identifiers) to choose the sample. Or the population can be numbered and a random number generator can choose the sample. Here, each member of the population has an equal chance of being chosen. If the goal of the technique is to get a very random sample, this is the best method to use. But it requires having a list of the whole population, which is not always realistic.

For example, suppose you want to take a random sample of university students. Each student is already numbered by their student ID. You could randomly select the members of your sample by using a random number generator to randomly select student ID numbers.

Stratified sampling and proportionate stratified sampling

If there are concerns that a random sample might not fully represent a population (e.g. one portion of the population is small compared to another), the best sampling technique to use is **stratified random sampling**. In this case, divide the population into groups called strata and then take a random sample from each stratum. The stratum are chosen to be a portion of the population that needs to be represented in the sample. Each stratum needs to be mutually exclusive from any other strata. That means that each member of the population can only belong to one stratum.

For example, you could stratify (group) your university population by faculty and then choose a simple random sample from each stratum (each faculty) to get a stratified random sample. As a student should only belong to one faculty, the groups are mutually exclusive. Further, this method ensures our sample is representative of the population by choosing students from each faculty at the university. Using the students per faculty example above, if the sample size is 100, to get a stratified sample, you would randomly select 20 students from each faculty (as there are 5 faculties and 100 students, choose an equal number from each faculty).

If the size of the sample is proportionate to the size of the strata, this is called **proportionate stratified random sampling**. If you wanted a proportionate stratified random sample for students by faculty, you would randomly select 20 students from business, 25 students from arts, 30 from science and engineering, 15 from nursing, and 10 from education (i.e. proportional to the number of students in each faculty). This technique is best used when there are large differences in the proportion of each group. For example, if the faculty of business had 50% of the students and the faculty of nursing only had 1% of the students, it would not be good to have an equal number of students from each faculty.

To randomly choose students from each faculty, a random sampling technique needs to be used. This could be simple random sampling or systematic random sampling (see below).

Cluster sampling
To choose a cluster sample, divide the population into clusters (groups) and then randomly select one of the clusters. That cluster is your sample. Further, the clusters need to be homogeneous and each cluster needs to be representative of the population. For example, suppose the university has a series of foundational classes that every student has to take and that students in these classes come from all faculties. Then we would randomly select one of these classes to be our sample. Again, to randomly select the four departments, you have to use a random sampling technique. Here, you could number all of the classes and then use a random number generator to choose one of them.

If one cluster is too small for the sample, you can choose more than one cluster. For example, if you want your sample to be 120 students but each of the foundational classes only have 30 students in them, you can randomly select 4 classes to get to your desired sample size.

Cluster sampling can be very convenient as the members of the sample are in one location. In the above example, the sample are in one class so you would just go to the one class and collect your sample. Notice that for stratified sampling, we would have to find each student chosen from each faculty. Thus, cluster sampling can save time and money. But it does present a real chance of undercoverage. If the foundational class chosen is at a time that nursing students are at a practicum, then that faculty would be undercovered. This means that cluster sampling can result in non-representative samples. This is only a good technique to use if the clusters are very similar to each other and each cluster would be representative of the population.

Cluster vs. stratified

Cluster sampling and stratified sampling are often confused. In each case, the population is divided into groups. But, in stratified sampling, a few people from all groups (strata) are chosen. While in cluster sampling, all of the people from a group (cluster) are chosen.

Additionally how the groups are chosen are different. In stratified sampling, the groups are chosen to be heterogeneous (i.e. each group has a different quality). As an example, breaking a university into different faculties results in groups that are heterogeneous as each group has a different quality (i.e. faculty) than the other groups. On the other hand, in cluster sampling, the groups are chosen to be homogeneous (i.e. the groups have similar qualities). That is, we want each cluster to be similar to the other groups.

Systematic random sampling

To choose a **systematic random sample**, randomly select a starting point and take every kth piece of data from a list of the population. For example, to choose a random sample of university students, you could use a list of all student names that are numbered by their student ID. Suppose there are 14,000 students at the university. To perform systematic random sampling, use a random number generator to pick a student ID number that represents the first name in the sample. Then calculate *k*. To do this, *k* is found by taking the population size (14,000) and dividing by the size of the sample (100). In this case, this results in 140. Thus, from your random starting point, choose every 140th name thereafter until you have a total of 100 names. If you reach the end of the list before completing your sample you simply go back to the beginning and keep going until the sample is complete.

Be careful: *k* needs to be large enough to ensure that you cycle through all the names. Otherwise the sample is not random nor is it representative. If *k* had been 10, then once the random starting point was chosen only 1000 names had a chance of being chosen which means that not everyone has an equal chance of being chosen. Further, depending on how the list is sorted, it may not be representative. For example, if our list of students is by faculty, then only certain faculties could make it in our sample. In our example, any *k* larger than 140 would be appropriate. Systematic sampling is frequently chosen because it is a simple method that can be easily implemented. But like simple random sampling, a list of the population is needed to do it properly.

There is a variation of systematic random sampling that can be used when the list of the population does not exist or is not available to the people doing the pull. For example, suppose you are doing a survey about people's satisfaction with a certain mall's hours. You won't have a list of all of the people who go to the mall. Instead, you may stand at an entrance to the mall and ask every fifth person who enters the mall to complete your survey. To ensure the sampling technique is representative, you'll want to do the survey multiple times at multiple locations. To ensure that the sampling technique is random, you'll want to randomly choose your starting times and locations. Having said that, this method would never be completely representative nor random. But may be your only choice if the population is not well defined.

Randomness and ethics

When we are performing a study, we cannot force people to be part of it. People have a right to say no and as researchers we need to seek informed consent. That is, the participants should know what they are being asked to do, how their information will be kept secure, if there are any risks to participation (and if so what they are), and how to see the results of the study. As such, people can choose not to participant in a study.

Thus, all studies involving humans are never completely random nor completely representative. Our goal when implementing sampling techniques is to minimize any bias that may come into the study because of this.

Convenience sampling

A type of sampling that is non-random is **convenience sampling**. Convenience sampling involves using results that are readily available. For example, a computer software store conducts a marketing study by interviewing potential customers who happen to be in the store browsing through the available software. The results of convenience sampling may be very good in some cases and highly biased (favour certain outcomes) in others. This is not a valid sampling technique when it comes to statistical inference. That is, if the data is collected using a convenience sample, then no conclusions can be made about the population from the sample.

With replacement or without replacement
True random sampling is done with **replacement**. That is, once a member is picked, that member goes back into the population and thus may be chosen more than once. However, for practical reasons, in most populations, simple random sampling is done without replacement. Surveys are typically done without replacement. That is, a member of the population may be chosen only once. Most samples are taken from large populations and the sample tends to be small in comparison to the population. Since this is the case, sampling without replacement is approximately the same as sampling with replacement because the chance of picking the same individual more than once with replacement is very low.

Too illustrate how small of chance it is, consider a university with a population of 10,000 people. Suppose you want to pick a sample of 1,000 randomly for a survey. **For any particular sample of 1,000**, if you are sampling **with replacement**,

- the chance of picking the first person is 1,000 out of 10,000 (0.1000);
- the chance of picking a different second person for this sample is 999 out of 10,000 (0.0999);
- the chance of picking the same person again is 1 out of 10,000 (very low).

If you are sampling **without replacement**,

- the chance of picking the first person for any particular sample is 1000 out of 10,000 (0.1000);
- the chance of picking a different second person is 999 out of 9,999 (0.0999);
- you do not replace the first person before picking the next person.

Compare the fractions 999/10,000 and 999/9,999. For accuracy, carry the decimal answers to four decimal places. To four decimal places, these numbers are equivalent (0.0999).

Sampling without replacement instead of sampling with replacement becomes a mathematical issue only when the population is small. For example, if the population is 25 people, the sample is ten, and you are sampling **with replacement for any particular sample**, then the chance of picking the first person is ten out of 25, and the chance of picking a different second person is nine out of 25 (you replace the first person).

If you sample **without replacement**, then the chance of picking the first person is ten out of 25, and then the chance of picking the second person (who is different) is nine out of 24 (you do not replace the first person).

Compare the fractions 9/25 and 9/24. To four decimal places, 9/25 = 0.3600 and 9/24 = 0.3750. To four decimal places, these numbers are not equivalent.

A study is done to determine the average tuition that San Jose State undergraduate students pay per semester. Each student in the following samples is asked how much tuition he or she paid for the Fall semester. What is the type of sampling in each case?

1. A sample of 100 undergraduate San Jose State students is taken by organizing the students' names by classification (freshman, sophomore, junior, or senior), and then selecting 25 students from each.
2. A random number generator is used to select a student from the alphabetical

listing of all undergraduate students in the Fall semester. Starting with that student, every 50th student is chosen until 75 students are included in the sample.

3. A completely random method is used to select 75 students. Each undergraduate student in the fall semester has the same probability of being chosen at any stage of the sampling process.
4. The freshman, sophomore, junior, and senior years are numbered one, two, three, and four, respectively. A random number generator is used to pick two of those years. All students in those two years are in the sample.
5. An administrative assistant is asked to stand in front of the library one Wednesday and to ask the first 100 undergraduate students he encounters what they paid for tuition the Fall semester. Those 100 students are the sample.

a. stratified; b. systematic; c. simple random; d. cluster; e. convenience

Determine the type of sampling used (simple random, stratified, systematic, cluster, or convenience).

1. A soccer coach selects six players from a group of boys aged eight to ten, seven players from a group of boys aged 11 to 12, and three players from a group of boys aged 13 to 14 to form a recreational soccer team.
2. A pollster interviews all human resource personnel in five different high tech companies.
3. A high school educational researcher interviews 50 high school female teachers and 50 high school male teachers.
4. A medical researcher interviews every third cancer patient from a list of cancer patients at a local hospital.
5. A high school counselor uses a computer to generate 50 random numbers and then picks students whose names correspond to the numbers.
6. A student interviews classmates in his algebra class to determine how many pairs of jeans a student owns, on the average.

a. stratified; b. cluster; c. stratified; d. systematic; e. simple random; f.convenience

Try It

Determine the type of sampling used (simple random, stratified, systematic, cluster, or convenience).

A high school principal polls 50 freshmen, 50 sophomores, 50 juniors, and 50 seniors regarding policy changes for after school activities.

stratified

If we were to examine two samples representing the same population, even if we used random sampling methods for the samples, they would not be exactly the same. Just as there is variation in data, there is variation in samples. As you become accustomed to sampling, the variability will begin to seem natural.

Suppose ABC College has 10,000 part-time students (the population). We are interested in the average amount of money a part-time student spends on books in the fall term. Asking all 10,000 students is an almost impossible task.

Suppose we take two different samples.
First, we use convenience sampling and survey ten students from a first term organic chemistry class. Many of these students are taking first term calculus in addition to the organic chemistry class. The amount of money they spend on books is as follows:
$128 $87 $173 $116 $130 $204 $147 $189 $93 $153
The second sample is taken using a list of senior citizens who take P.E. classes and taking every fifth senior citizen on the list, for a total of ten senior citizens. They spend:
$50 $40 $36 $15 $50 $100 $40 $53 $22 $22
It is unlikely that any student is in both samples.

a. Do you think that either of these samples is representative of (or is characteristic of) the entire 10,000 part-time student population?

a. No. The first sample probably consists of science-oriented students. Besides the chemistry course, some of them are also taking first-term calculus. Books for these classes tend to be expensive. Most of these students are, more than likely, paying more than the average part-time student for their books. The second sample is a group of senior citizens who are, more than likely, taking courses for health and interest. The amount of money they spend on books is probably much less than the average parttime student. Both samples are biased. Also, in both cases, not all students have a chance to be in either sample.

b. Since these samples are not representative of the entire population, is it wise to use the results to describe the entire population?

b. No. For these samples, each member of the population did not have an equally likely chance of being chosen.

Now, suppose we take a third sample. We choose ten different part-time students from the disciplines of chemistry, math, English, psychology, sociology, history, nursing, physical education, art, and early childhood development. (We assume that these are the only disciplines in which part-time students at ABC College are enrolled and that an equal number of part-time students are enrolled in each of the disciplines.) Each student is chosen using simple random sampling. Using a calculator, random numbers are generated and a student from a particular discipline is selected if he or she has a corresponding number. The students spend the following amounts:
$180 $50 $150 $85 $260 $75 $180 $200 $200 $150

c. Is the sample biased?

c. The sample is unbiased, but a larger sample would be recommended to increase the likelihood that the sample will be close to representative of the population. However, for a biased sampling technique, even a large sample runs the risk of not being representative of the population.

Students often ask if it is "good enough" to take a sample, instead of surveying the entire population. If the survey is done well, the answer is yes.

Try It

A local radio station has a fan base of 20,000 listeners. The station wants to know if its audience would prefer more music or more talk shows. Asking all 20,000 listeners is an almost impossible task.

The station uses convenience sampling and surveys the first 200 people they meet at one of the station’s music concert events. 24 people said they’d prefer more talk shows, and 176 people said they’d prefer more music.

Do you think that this sample is representative of (or is characteristic of) the entire 20,000 listener population?

Try It Solutions

The sample probably consists more of people who prefer music because it is a concert event. Also, the sample represents only those who showed up to the event earlier than the majority. The sample probably doesn’t represent the entire fan base and is probably biased towards people who would prefer music.

Variation in Data

Variation is present in any set of data. For example, 16-ounce cans of beverage may contain more or less than 16 ounces of liquid. In one study, eight 16 ounce cans were measured and produced the following amount (in ounces) of beverage:

15.8 16.1 15.2 14.8 15.8 15.9 16.0 15.5

Measurements of the amount of beverage in a 16-

ounce can may vary because different people make the measurements or because the exact amount, 16 ounces of liquid, was not put into the cans. Manufacturers regularly run tests to determine if the amount of beverage in a 16-ounce can falls within the desired range.

Be aware that as you take data, your data may vary somewhat from the data someone else is taking for the same purpose. This is completely natural. However, if two or more of you are taking the same data and get very different results, it is time for you and the others to reevaluate your data-taking methods and your accuracy.

Variation in Samples

It was mentioned previously that two or more **samples** from the same **population**, taken randomly, and having close to the same characteristics of the population will likely be different from each other. Suppose Doreen and Jung both decide to study the average amount of time students at their college sleep each night. Doreen and Jung each take samples of 500 students. Doreen uses systematic sampling and Jung uses cluster sampling. Doreen's sample will be different from Jung's sample. Even if Doreen and Jung used the same sampling method, in all likelihood their samples would be different. Neither would be wrong, however.

Think about what contributes to making Doreen's and Jung's samples different.

If Doreen and Jung took larger samples (i.e. the number of data values is increased), their sample results (the average amount of time a student sleeps) might be closer to the actual population average. But still, their samples would be, in all likelihood, different from each other. This **variability in samples** cannot be stressed enough.

Size of a Sample

The size of a sample (often called the number of observations) is important. The examples you have seen in this book so far have been small. Samples of only a few hundred observations, or even smaller, are sufficient for many purposes. In polling, samples that are from 1,200 to 1,500 observations are considered large enough and good enough if the survey is random and is well done. You will learn why when you study confidence intervals.

Be aware that many large samples are biased. For example, call-in surveys are invariably biased, because people choose to respond or not.

Critical Evaluation

We need to evaluate the statistical studies we read about critically and analyze them before accepting the results of the studies. We listed common problems with sampling techniques above. We re-iterate them here and add a few additional ones.

- Problems with samples: A sample must be representative of the population. A sample that is not representative of the population is biased. Biased samples that are not representative of the population give results that are inaccurate and not valid.
- Self-selected samples: Responses only by people who choose to respond, such as call-in surveys, are often unreliable.
- Sample size issues: Samples that are too small may be unreliable. Larger samples are better, if possible. In some situations, having small samples is unavoidable and can still be used to draw conclusions. Examples: crash testing cars or medical testing for rare conditions
- Undue influence: collecting data or asking questions in a way that influences the response
- Non-response or refusal of subject to participate: The collected responses may no longer be representative of the population. Often, people with strong positive or negative opinions may answer surveys, which can affect the results.

- Causality: A relationship between two variables does not mean that one causes the other to occur. They may be related (correlated) because of their relationship through a different variable.
- Self-funded or self-interest studies: A study performed by a person or organization in order to support their claim. Is the study impartial? Read the study carefully to evaluate the work. Do not automatically assume that the study is good, but do not automatically assume the study is bad either. Evaluate it on its merits and the work done.
- Misleading use of data: improperly displayed graphs, incomplete data, or lack of context
- Confounding: When the effects of multiple factors on a response cannot be separated. Confounding makes it difficult or impossible to draw valid conclusions about the effect of each factor.

References

Gallup-Healthways Well-Being Index. http://www.well-beingindex.com/default.asp (accessed May 1, 2013).

Gallup-Healthways Well-Being Index. http://www.well-beingindex.com/methodology.asp (accessed May 1, 2013).

Gallup-Healthways Well-Being Index. http://www.gallup.com/poll/146822/gallup-healthways-index-questions.aspx (accessed May 1, 2013).

Data from http://www.bookofodds.com/Relationships-Society/Articles/A0374-How-George-Gallup-Picked-the-President

Dominic Lusinchi, “’President’ Landon and the 1936 *Literary Digest* Poll: Were Automobile and Telephone Owners to Blame?” Social Science History 36, no. 1: 23-54 (2012), http://ssh.dukejournals.org/content/36/1/23.abstract (accessed May 1, 2013).

“The Literary Digest Poll,” Virtual Laboratories in Probability and Statistics http://www.math.uah.edu/stat/data/LiteraryDigest.html (accessed May 1, 2013).

“Gallup Presidential Election Trial-Heat Trends, 1936–2008,” Gallup Politics http://www.gallup.com/poll/110548/gallup-presidential-election-trialheat-trends-19362004.aspx#4 (accessed May 1, 2013).

The Data and Story Library, http://lib.stat.cmu.edu/DASL/Datafiles/USCrime.html (accessed May 1, 2013).

LBCC Distance Learning (DL) program data in

2010-2011, http://de.lbcc.edu/reports/2010-11/future/highlights.html#focus (accessed May 1, 2013).

Data from San Jose Mercury News

Chapter Review

Data are individual items of information that come from a population or sample. Data may be classified as categorical nominal, categorical ordinal, quantitative continuous, or quantitative discrete.

Because it is not practical to measure the entire population in a study, researchers use samples to represent the population. A random sample is a representative group from the population chosen by using a method that gives each individual in the population an equal chance of being included in the sample. Random sampling methods include simple random sampling, stratified sampling, cluster sampling, and systematic sampling. Convenience sampling is a nonrandom method of choosing a sample that often produces biased data.

Samples that contain different individuals result in different data. This is true even when the samples are well-chosen and representative of the population. When properly selected, larger samples model the population more closely than smaller samples. There are many different potential problems that can affect the reliability of a sample. Statistical data needs to be critically analyzed, not simply accepted.

HOMEWORK

For the following exercises, identify the type of data that would be used to describe a response (quantitative discrete, quantitative continuous, or categorical), and give an example of the data.

number of tickets sold to a concert

---

quantitative discrete, 150

percent of body fat

---

quantitative continuous, 19.2%

favorite baseball team

---

categorical, Oakland A’s

time in line to buy groceries

quantitative continuous, 7.2 minutes

number of students enrolled at Evergreen Valley College

quantitative discrete, 11,234 students

most-watched television show

categorical, Dancing with the Stars

brand of toothpaste

categorical, Crest

distance to the closest movie theatre

quantitative continuous, 8.32 miles

age of executives in Fortune 500 companies quantitative continuous, 47.3 years

Use the following information to answer the next two exercises: A study was done to determine the age, number of times per week, and the duration (amount of time) of resident use of a local park in Vancouver. The first house in the neighbourhood around the park was selected randomly and then every 8th house in the neighbourhood around the park was interviewed.

"Number of times per week" is what type of data?

1. nominal categorical ordinal
2. quantitative discrete
3. quantitative continuous
4. categorical nominal
5. categorical ordinal

b

"Duration (amount of time)" is what type of data?

1. categorical discrete
2. quantitative discrete
3. quantitative continuous
4. categorical nominal

5. categorical ordinal

---

c

Airline companies are interested in the consistency of the number of babies on each flight, so that they have adequate safety equipment. Suppose an airline conducts a survey. Over Thanksgiving weekend, it surveys six flights from Montreal to Halifax to determine the number of babies on the flights. It determines the amount of safety equipment needed by the result of that study.

1. Using complete sentences, list three things wrong with the way the survey was conducted.
2. Using complete sentences, list three ways that you would improve the survey if it were to be repeated.

---

1. The survey was conducted using six similar flights.
   The survey would not be a true representation of the entire population of air travelers.
   Conducting the survey on a holiday weekend will not produce representative results.

2. Conduct the survey during different times of the year.
   Conduct the survey using flights to and from various locations.
   Conduct the survey on different days of the week.

Suppose you want to determine the mean number of cans of soda drunk each month by students in their twenties at your school. Describe a possible sampling method in three to five complete sentences. Make the description detailed.

Answers will vary. Sample Answer: You could use a systematic sampling method. Stop the tenth person as they leave one of the buildings on campus at 9:50 in the morning. Then stop the tenth person as they leave a different building on campus at 1:50 in the afternoon.

Name the sampling method used in each of the following situations:

1. A woman in the airport is handing out questionnaires to travelers asking them to evaluate the airport's service. She does not ask travelers who are hurrying through the

airport with their hands full of luggage, but instead asks all travelers who are sitting near gates and not taking naps while they wait.

2. A teacher wants to know if her students are doing homework, so she randomly selects rows two and five and then calls on all students in row two and all students in row five to present the solutions to homework problems to the class.
3. The marketing manager for an electronics chain store wants information about the ages of its customers. Over the next two weeks, at each store location, 100 randomly selected customers are given questionnaires to fill out asking for information about age, as well as about other variables of interest.
4. The librarian at a public library wants to determine what proportion of the library users are children. The librarian has a tally sheet on which she marks whether books are checked out by an adult or a child. She records this data for every fourth patron who checks out books.
5. A political party wants to know the reaction of voters to a debate between the candidates. The day after the debate, the party's polling staff calls 1,200 randomly selected phone numbers. If a registered voter answers the phone or is available to

come to the phone, that registered voter is asked whom he or she intends to vote for and whether the debate changed his or her opinion of the candidates.

---

convenience cluster stratified systematic simple random

In advance of the 1936 Presidential Election, a magazine titled Literary Digest released the results of an opinion poll predicting that the republican candidate Alf Landon would win by a large margin. The magazine sent post cards to approximately 10,000,000 prospective voters. These prospective voters were selected from the subscription list of the magazine, from automobile registration lists, from phone lists, and from club membership lists. Approximately 2,300,000 people returned the postcards.

1. Think about the state of the United States in 1936. Explain why a sample chosen from magazine subscription lists, automobile registration lists, phone books, and club membership lists was not representative of the population of the United States at that time.
2. What effect does the low response rate have on the reliability of the sample?
3. Are these problems examples of sampling

error or nonsampling error?

4. During the same year, George Gallup conducted his own poll of 30,000 prospective voters. His researchers used a method they called "quota sampling" to obtain survey answers from specific subsets of the population. Quota sampling is an example of which sampling method described in this module?

---

1. The country was in the middle of the Great Depression and many people could not afford these “luxury” items and therefore not able to be included in the survey.
2. Samples that are too small can lead to sampling bias.
3. sampling error
4. stratified

YouPolls is a website that allows anyone to create and respond to polls. One question posted April 15 asks:

“Do you feel happy paying your taxes when members of the Obama administration are allowed to ignore their tax liabilities?”[footnote]

lastbaldeagle. 2013. On Tax Day, House to Call for Firing Federal Workers Who Owe Back

Taxes. Opinion poll posted online at: http://www.youpolls.com/details.aspx?id=12328 (accessed May 1, 2013).

As of April 25, 11 people responded to this question. Each participant answered “NO!”

Which of the potential problems with samples discussed in this module could explain this connection?

Self-Selected Samples: Only people who are interested in the topic are choosing to respond. Sample Size Issues: A sample with only 11 participants will not accurately represent the opinions of a nation.

Undue Influence: The question is wording in a specific way to generate a specific response. Self-Funded or Self-Interest Studies: This question was generated to support one person’s claim and it was designed to get the answer that the person desires.

Glossary

Cluster Sampling

a method for selecting a random sample and dividing the population into groups (clusters); use simple random sampling to select a set of clusters. Every individual in the chosen clusters is included in the sample.

Continuous Random Variable

a random variable (RV) whose outcomes are measured; the height of trees in the forest is a continuous RV.

Convenience Sampling

a nonrandom method of selecting a sample; this method selects individuals that are easily accessible and may result in biased data.

Discrete Random Variable

a random variable (RV) whose outcomes are counted

Nonsampling Error

an issue that affects the reliability of sampling data other than natural variation; it includes a variety of human errors including poor study design, biased sampling methods, inaccurate information provided by study participants, data entry errors, and poor analysis.

Qualitative Data

See Data.

Quantitative Data

See Data.

Random Sampling

a method of selecting a sample that gives every member of the population an equal chance of being selected.

Sampling Bias
not all members of the population are equally likely to be selected

Sampling Error
the natural variation that results from selecting a sample to represent a larger population; this variation decreases as the sample size increases, so selecting larger samples reduces sampling error.

Sampling with Replacement
Once a member of the population is selected for inclusion in a sample, that member is returned to the population for the selection of the next individual.

Sampling without Replacement
A member of the population may be chosen for inclusion in a sample only once. If chosen, the member is not returned to the population before the next selection.

Simple Random Sampling
a straightforward method for selecting a random sample; give each member of the population a number. Use a random number generator to select a set of labels. These

randomly selected labels identify the members of your sample.

Stratified Sampling

a method for selecting a random sample used to ensure that subgroups of the population are represented adequately; divide the population into groups (strata). Use simple random sampling to identify a proportionate number of individuals from each stratum.

Systematic Sampling

a method for selecting a random sample; list the members of the population. Use simple random sampling to select a starting point in the population. Let k = (number of individuals in the population)/(number of individuals needed in the sample). Choose every kth individual in the list starting with the one that was randomly selected. If necessary, return to the beginning of the population list to complete your sample.

Experimental Design and Ethics

Does aspirin reduce the risk of heart attacks? Is one brand of fertilizer more effective at growing roses than another? Is fatigue as dangerous to a driver as the influence of alcohol? Questions like these are answered using randomized experiments. In this module, you will learn important aspects of experimental design. Proper study design ensures the production of reliable, accurate data.

The purpose of an experiment is to investigate the relationship between two variables. When one variable causes change in another, we call the first variable the **independent variable** or **explanatory variable**. The affected variable is called the **dependent variable** or **response variable**. In a randomized experiment, the researcher manipulates values of the explanatory variable and measures the resulting changes in the response variable. The different values of the explanatory variable are called **treatments**. An **experimental unit** is a single object or individual to be measured.

You want to investigate the effectiveness of vitamin E in preventing disease. You recruit a group of subjects and ask them if they regularly take vitamin E. You notice that the subjects who take vitamin E exhibit better health on average than those who do not. Does this prove that vitamin E is effective in disease prevention? It does not. There are many differences between the two groups compared in addition to vitamin E consumption. People who take vitamin E regularly often take other steps to improve their health: exercise, diet, other vitamin supplements, choosing not to smoke. Any one of these factors could be influencing health. As described, this study does not prove that vitamin E is the key to disease prevention.

Additional variables that can cloud a study are called **lurking variables**. In order to prove that the explanatory variable is causing a change in the response variable, it is necessary to isolate the explanatory variable. The researcher must design her experiment in such a way that there is only one difference between groups being compared: the planned treatments. This is accomplished by the **random assignment** of experimental units to treatment groups. When subjects are assigned treatments randomly, all of the potential lurking variables are spread equally among the groups. At this point the only difference between groups is the one imposed by the researcher. Different outcomes measured in the response variable, therefore, must be a direct result of the different treatments. In this way, an experiment can prove a cause-and-effect connection between the explanatory and response variables.

The power of suggestion can have an important influence on the outcome of an experiment. Studies have shown that the expectation of the study participant can be as important as the actual medication. In one study of performance-enhancing drugs, researchers noted:

> *Results showed that believing one had taken the substance resulted in [*performance*] times almost as fast as those associated with consuming the drug itself. In contrast, taking the drug without knowledge yielded no significant performance increment.* [footnote] McClung, M. Collins, D. “Because I know it will!”: placebo effects of an ergogenic aid on athletic performance. Journal of Sport & Exercise Psychology. 2007 Jun. 29(3):382-94. Web. April 30, 2013.

When participation in a study prompts a physical response from a participant, it is difficult to isolate the effects of the explanatory variable. To counter the power of suggestion, researchers set aside one treatment group as a **control group**. This group is given a **placebo** treatment–a treatment that cannot influence the response variable. The control group helps researchers balance the effects of being in an experiment with the effects of the active treatments. Of course, if you are participating in a study and you know that you are receiving a pill which contains no actual medication, then the power of suggestion is no longer a factor. **Blinding** in a randomized experiment preserves the power of suggestion. When a person involved in a research study is blinded, he does not know who is receiving the active treatment(s) and who is receiving the placebo treatment. A **double-blind experiment** is one in which both the subjects and the researchers involved with the subjects are blinded.

The Smell & Taste Treatment and Research Foundation conducted a study to investigate whether smell can affect learning. Subjects completed mazes multiple times while wearing masks. They completed the pencil and paper mazes three times wearing floral-scented masks, and three times with unscented masks. Participants were assigned at random to wear the floral mask during the first three trials or during the last three trials. For each trial, researchers recorded the time it took to complete the maze and the subject's impression of the mask's scent: positive, negative, or neutral.

1. Describe the explanatory and response variables in this study.
2. What are the treatments?
3. Identify any lurking variables that could interfere with this study.
4. Is it possible to use blinding in this study?

1. The explanatory variable is scent, and the response variable is the time it takes to complete the maze.
2. There are two treatments: a floral-scented mask and an unscented mask.
3. All subjects experienced both treatments. The order of treatments was randomly assigned so there were no differences between the treatment groups. Random assignment eliminates the problem of lurking variables.
4. Subjects will clearly know whether they can smell flowers or not, so subjects cannot be blinded in this study. Researchers timing the mazes can be blinded, though. The researcher who is observing a subject will not know which mask is being worn.

References

“Vitamin E and Health,” Nutrition Source, Harvard School of Public Health, http://www.hsph.harvard.edu/nutritionsource/vitamin-e/ (accessed May 1, 2013).

Stan Reents. “Don’t Underestimate the Power of Suggestion,” athleteinme.com, http://www.athleteinme.com/ArticleView.aspx?id=1053 (accessed May 1, 2013).

Ankita Mehta. “Daily Dose of Aspiring Helps Reduce Heart Attacks: Study,” International Business Times, July 21, 2011. Also available online at http://www.ibtimes.com/daily-dose-aspirin-helps-reduce-heart-attacks-study-300443 (accessed May 1, 2013).

The Data and Story Library, http://lib.stat.cmu.edu/DASL/Stories/ScentsandLearning.html (accessed May 1, 2013).

M.L. Jacskon et al., “Cognitive Components of Simulated Driving Performance: Sleep Loss effect and Predictors,” Accident Analysis and Prevention Journal, Jan no. 50 (2013), http://www.ncbi.nlm.nih.gov/pubmed/22721550 (accessed May 1, 2013).

“Earthquake Information by Year,” U.S. Geological Survey. http://earthquake.usgs.gov/earthquakes/eqarchives/year/ (accessed May 1, 2013).

“Fatality Analysis Report Systems (FARS) Encyclopedia,” National Highway Traffic and Safety Administration. http://www-fars.nhtsa.dot.gov/Main/index.aspx (accessed May 1, 2013).

Data from www.businessweek.com (accessed May 1, 2013).

Data from www.forbes.com (accessed May 1, 2013).

"America's Best Small Companies," http://www.forbes.com/best-small-companies/list/ (accessed May 1, 2013).

U.S. Department of Health and Human Services, Code of Federal Regulations Title 45 Public Welfare Department of Health and Human Services Part 46 Protection of Human Subjects revised January 15, 2009. Section 46.111:Criteria for IRB Approval of Research.

"April 2013 Air Travel Consumer Report," U.S. Department of Transportation, April 11 (2013), http://www.dot.gov/airconsumer/april-2013-air-travel-consumer-report (accessed May 1, 2013).

Lori Alden, "Statistics can be Misleading," econoclass.com, http://www.econoclass.com/misleadingstats.html (accessed May 1, 2013).

Maria de los A. Medina, "Ethics in Statistics," Based on "Building an Ethics Module for Business, Science, and Engineering Students" by Jose A. Cruz-Cruz and William Frey, Connexions, http://cnx.org/content/m15555/latest/ (accessed May 1, 2013).
Andrew Gelman, "Open Data and Open Methods," Ethics and Statistics, http://www.stat.columbia.edu/~gelman/research/

published/ChanceEthics1.pdf (accessed May 1, 2013).

Chapter Review

A poorly designed study will not produce reliable data. There are certain key components that must be included in every experiment. To eliminate lurking variables, subjects must be assigned randomly to different treatment groups. One of the groups must act as a control group, demonstrating what happens when the active treatment is not applied. Participants in the control group receive a placebo treatment that looks exactly like the active treatments but cannot influence the response variable. To preserve the integrity of the placebo, both researchers and subjects may be blinded. When a study is designed properly, the only difference between treatment groups is the one imposed by the researcher. Therefore, when groups respond differently to different treatments, the difference must be due to the influence of the explanatory variable.

“An ethics problem arises when you are considering an action that benefits you or some cause you support, hurts or reduces benefits to others, and violates some rule.”[footnote] Ethical violations in statistics are not always easy to spot. Professional associations and federal agencies post guidelines for proper conduct. It is important that you learn basic statistical procedures so that you can recognize proper data analysis.

Glossary

Explanatory Variable

the **independent variable** in an experiment; the value controlled by researchers

Treatments

different values or components of the explanatory variable applied in an experiment

Response Variable

the **dependent variable** in an experiment; the value that is measured for change at the end of an experiment

Experimental Unit

any individual or object to be measured

Lurking Variable

a variable that has an effect on a study even though it is neither an explanatory variable nor a response variable

Random Assignment

the act of organizing experimental units into treatment groups using random methods

Control Group

a group in a randomized experiment that

receives an inactive treatment but is otherwise managed exactly as the other groups

Informed Consent

Any human subject in a research study must be cognizant of any risks or costs associated with the study. The subject has the right to know the nature of the treatments included in the study, their potential risks, and their potential benefits. Consent must be given freely by an informed, fit participant.

Institutional Review Board

a committee tasked with oversight of research programs that involve human subjects

Placebo

an inactive treatment that has no real effect on the explanatory variable

Blinding

not telling participants which treatment a subject is receiving

Double-blinding

the act of blinding both the subjects of an experiment and the researchers who work with the subjects

Introduction to Descriptive Statistics

class="introduction" When you have large amounts of data, you will need to organize it in a way that makes sense. These ballots from an election are rolled together with similar ballots to keep them organized. (credit: William Greeson)

Chapter objective

By the end of this chapter, the student should be able to:

- Display data graphically and interpret graphs: pie charts, bar graphs, histograms and box plots.
- Recognize, describe, calculate, and interpret measures of location: quartiles and percentiles.

- Recognize, describe, calculate, and interpret measures of centre: mean, median and mode.
- Recognize, describe, calculate, and interpret measures of variation: variance, standard deviation, range, interquartile range and coefficient of variation.

Once you have collected data, what will you do with it? Data can be described and presented in many different formats. For example, suppose you are interested in buying a house in a particular area. You may have no clue about the house prices, so you might ask your real estate agent to give you a sample data set of prices. Looking at all the prices in the sample often is overwhelming. A better way might be to look at the median price and the variation of prices. The median and variation are just two ways that you will learn to describe data. Your agent might also provide you with a graph of the data.

In this chapter, you will study visual and numerical ways to describe and display your data. This area of statistics is called **Descriptive Statistics.** If you have collected 200 data values, just looking at them won’t tell anyone much about the data. Instead, you want to summarize the raw data in a way that you can better understand what’s going on.

Categorical data is summarized usually using a visual representation like a pie chart or a bar graph. The numerical summary for categorical data would be a percentage, fraction or decimal.

For quantitative data, it is a bit more involved. In general, there are three components to a good summary of quantitative data: a visual representation, a measure of centre, and a measure of variation.

The visual representation can give you a sense of the centre and variation in the data, but is very useful for determining the **shape** of the data. Is the data all clustered together? Are there a bunch of data on one side, but a few on the other? Do all of the data values occur with the same frequency? The shape describes this. Histograms and box plots are both visual representations of quantitative data.

Measures of centre, also known as **averages** or **measures of central tendency**, provide a value(s) that gives us a sense of a typical value in the data set. This doesn’t tell us about a specific member of the population, but instead lets us know what the average one is like. Measures of centre we will learn about include the mean, median, and mode.

Though a measure of centre tells us about a typical value in a data set, **measures of variation** tell us how much the data values vary from each other. Are they all clumped together? Are they all spread out? Measures of variation can tell us how consistent or how volatile the data is. If we are analyzing stock prices, the more variation there is then the more volatile and risky the investment is. But the rewards may be greater! Measures of variation that we will learn about include range, variance, standard deviation, interquartile range, and the coefficient of variation.

When we describe the shape, centre, and variation of the data, we are describing the **distribution** of the data. If we only focus on one aspect of the distribution (say the centre), then we miss out on some important information, which is why we always want to consider all three aspects when summarizing quantitative data. For example, suppose two stock prices have the same average price. If we only look at the average, we might think they are equivalent. But if one of them has greater variation, then that means that one is more volatile and riskier than the other one.

Box plots (or box and whisker diagrams) are a special type of visual representation that includes both visual and numerical elements. A box plot divides the data into quarters (or **quartiles**). Thus, a box plot contains a measure of centre (the second quartile is the halfway point, called the median) and a measure of variation (the distance between the first quartile and the third quartile is called the interquartile range). The box plot can also give a sense of the data's shape. The box plot then is the only representation that we will see that gives us a sense of the distribution all in one representation (i.e. gives a sense of centre, variation, and distribution). It also has an additional benefit of identifying outliers. **Outliers** are data values that are abnormal. That is, they differ significantly from the other data values. A box plot shows if there are any outliers.

This chapter will go over descriptive statistics by focusing on visual and numerical representations of data. Though categorical data is discussed, the main focus will be on determining the distribution and outliers for quantitative data.

The vast majority of the time when conducting statistical studies, we will only have access to sample data. In this situation, we will want to analyze the sample data to see if we can come to any conclusions about the population data. Once we make the leap from simply describing a sample to using that sample to draw conclusions about the population, we are doing **inferential statistics**. These concepts and techniques are covered in chapter seven and eight.

Key Idea

The distribution of sample data ideally mimics the distribution of the population. But the smaller the sample size the greater the potential for there to be differences between the two distributions. This means that, for a large enough sample size, the distribution of the sample generally gives a good idea of distribution of the population. This is an example of the **law of large numbers**. In other words, if the sample size is large enough and the data is collected properly, then the sample mean will most likely be a good estimate of the population mean, the sample standard deviation will most likely be a good estimate of the population standard deviation, and the shape of the sample data will most likely be a good estimate of the shape of the population.

Numerical Summaries

By the end of this section, we want to be able to describe the distribution of quantitative data (i.e. shape, centre and variation). In the previous section, we looked at the shape of quantitative data. This section focuses on numerical summaries of data for quantitative data. In particular, it focuses on measures of centre and measures of variation.

There are other numerical summaries of data called measures of location, which will be discussed in the next section.
To find the mean of this data, we need to find the number that balances the data equally on both sides. To find the mean of this data, we need to find the number that balances the data equally on both sides. Notice that the mean here is not a data value.

Measures of centre

Measures of centre or average give us a sense of what a typical value in a data set is. For example, the average number of children in a family in Canada is 1.9. This means that a typical family will have about 1.9 children. Obviously, no family has exactly 1.9 children, but this gives a sense of how many children families have on average. Further, some families may have 8 children. Others may have no children. The measure of centre gives a sense of what is going on in the middle of the data set.

> Even though you may wish to round an average to a whole number (especially when it is about the number of people), this is not necessary nor is it appropriate as it is giving a sense of the centre of the data, which is not necessarily an actual data value.

The "center" of a data set is a way of describing a typical value in a data set. The three most widely used measures of the "center" of the data are the **mean**, **median** and **mode**.

To explain these three measures of centre, let’s look at an example. Suppose we want to find the average weight of 50 people. To calculate the mean weight of the 50 people, we would add the 50 weights together and divide by 50. To find the median weight of the 50 people, order the data from least heavy to most heavy, and find the weight that splits the data into two equal parts. The mode is the most commonly occurring value. To find the mode, find the weight that occurs the most frequently.

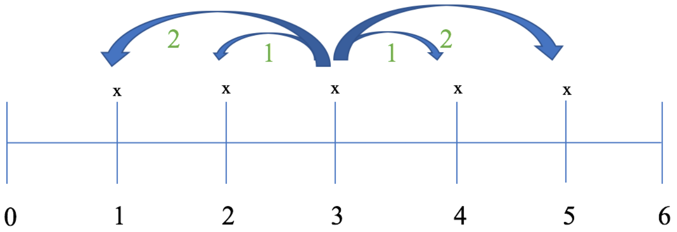
This section provides more details on how to find the measures of centre, the notation for the measures, and when it is best to used which measure.

NOTE

Though the words “mean” and “average” are sometimes used interchangeably, they do not necessarily mean the same thing. In general, “average” is any measure of centre and “mean” is a specific type of centre. Many people use average and mean as the same, but not always. For example, when people talk about average housing price, they are usually referring to the median house price.

Mean

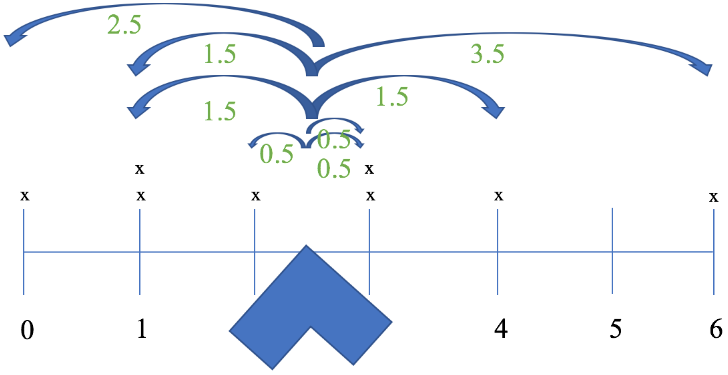
The mean of a data set can be thought of as a balancing point (or fulcrum). If you think of numbers as weighted, then the mean is the number that will balance the data values evenly. Suppose your data values are 1, 2, 3, 4, 5. Then the number that balances the data is 3. To go a little deeper, the balance point is three because the distance between 3 and the data values less than it is equal to the distance between 3 and the data values greater than it as shown in [link].



Let's try a harder example. Suppose our data values are 0, 1, 1, 2, 3, 3, 4, 6. The mean will be the number such that the total distance to the data values below it and the total distance to the data values above it are the same. Let's see 3 is the mean again. Then the distance between our suggested "mean" and 0 is 3; the distance between our "mean" and 1 is 2 (but there are two of them); and the distance between our "mean" and 2 is 1. That is, the distance between our "mean" and all of the data values below it are $3+2+2+1 = 8$. If 3 is actually our mean, then the total distance between 3 and the data values above it will also be 8. Let's check. The distance between our "mean" and 4 is 1; the distance between our "mean" and 6 is 3. The total distance above 3 is only 4. Therefore, 3 cannot be our mean as it doesn't balance our data.

The two data values of 3 were ignored as their distance from the suggested mean is 0. Therefore, they would not change the answer if included.

From our calculations above, the choice of 3 was too big as the lower was too heavy. Let's try 2.5 as our mean. If the mean is 2.5, then the distance between our "mean" and 0 is 2.5; the distance between our "mean" and 1 is 1.5 (but there are two of them); the distance between our "mean" and 2 is 0.5. Thus the total distance between our mean of 2.5 and the data values below is is 2.5 + 1.5 + 1.5 + 0.5 = 6. If 2.5 is our mean, then the total distance above 2.5 should also be 6. The distance between our "mean" and 3 is 0.5 (but there are two of them); the distance between our "mean" and 4 is 1.5; the distance between our "mean" and 6 is 3.5. Thus the total distance between the data values and our suggested mean of 2.5 is 0.5 + 0.5 + 1.5 + 3.5 = 6! Therefore, 2.5 is the mean for this data.



Thankfully we don't have to do these in-depth calculations and guesses each time. Instead the formula is pretty straight-forward.

The Greek letter μ (pronounced "mew") represents the **population mean**. That is, it is the mean for the population data.

Formula for Population Mean

$\mu = 1N\ \Sigma\ i=1\ N\ xi$

The letter used to represent the **sample mean** is an *x* with a bar over it (pronounced "*x* bar"): $\bar{x}$. It is the mean of a sample of data from the population.

The sample mean is an estimate of the population mean. One of the requirements for the **sample mean** to be a good estimate of the **population mean** is for the sample taken to be truly random.

Formula for Sample Mean

$x - = 1n\ \Sigma\ i=1\ n\ xi$

To see how the formula words, consider the sample: 1; 1; 1; 2; 2; 3; 4; 4; 4; 4; 4

$x - = 1 + 1 + 1 + 2 + 2 + 3 + 4 + 4 + 4 + 4 + 4\ 11 = 2.7$

Note: Since it is sample data, we use the symbol $\bar{x}$.

Application of the law of large numbers

If the size of a random sample is increased, then the sample mean will more likely be a better estimate of the population mean.

Note: Just because the sample size increases does not mean that the sample mean for the larger sample must be a better estimate. It is only that it is more likely to be a better estimate.

Median

On a road, the median is in the middle of the road. In statistics, the median is the middle data value (when the data is in order).

You can quickly find the *location* or position of the median by using the expression n + 1 2 .

The letter *n* is the total number of data values in the sample. If *n* is an odd number, the median is the middle value of the ordered data (ordered smallest to largest). If *n* is an even number, the median is equal to the two middle values added together and divided by two after the data has been ordered. For example, if the total number of data values is 97, then n + 1 2 = 97 + 1 2 = 49. The median is the 49th value in the ordered data. If the total number of data values is 100, then n + 1 2 = 100 + 1 2 = 50.5. The median occurs midway between the 50th and 51st values. The location of the median and the value of the median are **not** the same. The upper case letter *M* is often used to represent the median. The next example illustrates the location of the median and the value of the median.

Mode

Another measure of the center is the mode. The **mode** is the data value that occurs most frequently and at least twice.

A data set can have either

- no mode.
- one mode (unimodal)
- two modes (bimodal)
- or many modes (multimodal).

Consider the statistics exam scores for 20 students:
5053595963637272727272767881838484849093
The most frequent score is 72, which occurs five times. Mode = 72.

Note
The mode can be calculated for qualitative data as well as for quantitative data. For example, if the data set is: red, red, red, green, green, yellow, purple, black, blue, the mode is red.

AIDS data indicating the number of months a patient with AIDS lives after taking a new

antibody drug are as follows (smallest to largest):
3; 4; 8; 8; 10; 11; 12; 13; 14; 15; 15; 16; 16; 17; 17; 18; 21; 22; 22; 24; 24; 25; 26; 26; 27; 27; 29; 29; 31; 32; 33; 33; 34; 34; 35; 37; 40; 44; 44; 47;
Calculate the mean, median and mode.

The calculation for the mean is:

x – = [3 + 4 + (8) (2) + 10 + 11 + 12 + 13 + 14 + (15) (2) + (16) (2) + ... + 35 + 37 + 40 + (44) (2) + 47] 40 = 23.6
To find the median, *M*, first use the formula for the location. The location is:
n + 1 2 = 40 + 1 2 = 20.5
Starting at the smallest value, the median is located between the 20th and 21st values (the two 24s):
3; 4; 8; 8; 10; 11; 12; 13; 14; 15; 15; 16; 16; 17; 17; 18; 21; 22; 22; 24; 24; 25; 26; 26; 27; 27; 29; 29; 31; 32; 33; 33; 34; 34; 35; 37; 40; 44; 44; 47;

M = 24 + 24 2 = 24 To find the mode, we first have to determine if any data values repeat. If no data values repeat, there is no mode. Since 8 repeats, we know there is a mode. 8 repeats twice. We need to check if any data value repeats more than twice. If a data

value repeats more than twice, then it is the mode. Since no data value repeats more than twice, any data value that repeats twice is the mode.
Therefore, the modes are 8, 15, 16, 17, 22, 24, 26, 27, 29, 34, 44. This data set is multi-modal.

Suppose that in a small town of 50 people, one person earns $5,000,000 per year and the other 49 each earn $30,000. Which is the better measure of the "center": the mean, the median or the mode?

$x - = 5,000,000+49(30,000)\ 50\ =129,400$

$M = 30,000$

(There are 49 people who earn $30,000 and one person who earns $5,000,000.)

The mode is 30,000 as this data value occurs 49 times.

Since the median and mode are equal, lets focus on the median. The median is a better measure of the "center" than the mean because

49 of the values are 30,000 and one is 5,000,000. The 5,000,000 is an outlier. The 30,000 gives us a better sense of the middle of the data.

The above example highlights two important ideas:

- **Outliers**: We have defined outliers as data values that are significantly different from other data values, but we have not provided a way of finding them. This will be discussed in the next section. Regardless, we can see that 5 million is significantly different than 30 thousand in the above example.
- **Skew**: When a data set has outliers, the outliers have the potential to skew the mean. In the above example, the centre of the data is 30,000, but the mean is 129,400. Thus the outlier of 5 million is pulling the mean up. That is, it is skewing the centre value by pulling it to the right on the number line.

Comparing measures of centre

Above we have described how to find each of the measures of centre. But how do you choose which measure of centre to use in which situation? One option is to provide all three measures of centre, but sometimes this can be overwhelming to the audience. Instead you want to pick the best one that best describes that data. The following are some general guidelines for choosing the best measure of centre.

The mean is often the best measure of centre to use because it is the most well-known and familiar of the measures of centre. It is also the only measure of centre that is computed using all of the sample values. But the mean is susceptible to outliers. As was seen in [link], if there is an outlier, the mean can be pulled in one direction away from the centre.

Outliers are any data value that are significantly different from the other data values. In [link], the outlier is 5 million as it is significantly higher than the other data values. We will discuss how to find outliers in the section 2.3 (Boxplots).

If there is an outlier in the data set that is skewing the mean, the best measure of centre to use is the median as it is not susceptible to outliers.

But be careful: The presence of outliers does not necessarily mean that the median is the best measure of centre. Here are a couple of examples where this is the case:

1. Suppose there are 200 data values in a sample and one data value is an outlier, then the mean

will most likely not be affected by the outlier.
2. Suppose there is a data set that has outliers, but one is a high outlier and one is a low outlier. Then the outliers may balance out and not affect the mean.

The mode is best used for categorical data, but can sometimes be used for quantitative data. For example, in [link], the mode would be a good measure of centre because the majority of data values are the same.

In [link], since there are no outliers, the mean is the best measure of centre to use. In [link], since there is an outlier (5 million) and the mean and median are quite different, the median is the best measure of centre to use.

The following tables compare the measures of centre.

Measure	How Common
Mean	most familiar
Median	commonly used
Mode	sometimes used

Measure	Every Score Used
Mean	yes
Median	no
Mode	no

Measure	Affected by Outliers
Mean	yes
Median	no
Mode	no

How to mislead with averages

Consider the following situation: As you arrive at an open house in your preferred new home location, a neighbour comes up to you while he is walking his dog. “This is a great neighbourhood to live in! The average income in this neighbourhood is $60,000,” he tells you. You are pleased to hear how affluent the community is. A year after you’ve moved into your new home, the same neighbour comes to your door and asks you to sign a petition. “The city is overvaluing the homes in our neighbourhood again, which means more taxes. The average income in this neighbourhood is $20,000. We can’t afford these increases.” You dutifully sign the petition because you don’t want to pay more taxes, but you’re also confused. Wasn’t the average income a lot higher last year? What happened? Is your neighbour a liar? In this example, there are many different possible scenarios that could explain the discrepancy. But no matter what the scenario is, the neighbour is picking his statistics to fit his situation.

One scenario: The neighbour may be picking and choosing which measure of centre to use. Suppose that most people in the neighbourhood make around $20,000 a year, but there are a few people who live on the street with the super nice view who make $300,000 a year. Then in the first case, when he says the average income is $60,000, he has used the mean which has been pulled higher by the outliers of $300,000. He chose to use the mean to make the neighbourhood look more affluent than it really is.

But when he wanted to make the argument that the neighbourhood wasn’t as affluent and should be in a lower tax bracket, he changed which measure of centre to use. Instead he may have the used the median or mode because they aren’t influenced by the outliers.

Another scenario: The neighbour may be choosing how he defines income to help make his point. In the first case, he may have only used those who are employed to come up with the average salary. While in the second case, he may have used all adults in the neighbourhood including students living with their parents, stay-at-home parents, retired people or people out of work. Their incomes may be very low or non-existent which would skew the average to being lower. In this scenario, he may be using the same measure of centre, but is picking what he means by income to get the results he wants.

There are other possible scenarios. Can you think of any?

Skew

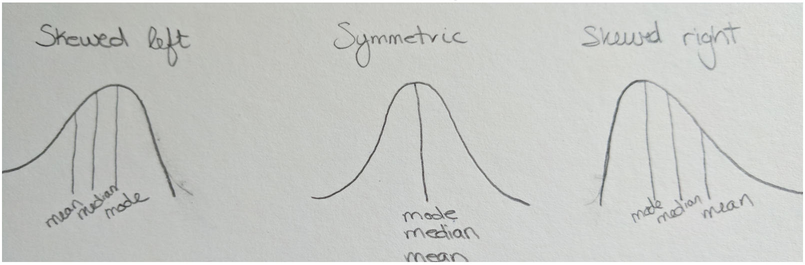
As has been noted above, if there are outliers in a data set, this can cause the mean to be pulled up or down (i.e. be either higher than expected or lower than expected) by these outliers. Outliers don't have to be present for this to happen. Essentially, any time that there are data values that cause the mean and median to be significantly different, then we say the data is **skewed**.

- If the mean is significantly larger than the median and the histogram has a long tail on the right, then the data is right skewed or positively skewed.
- If the mean is significantly smaller than the median and the histogram has a long tail on the left, then the data is left skewed or negatively skewed.
- If the mean and the median are approximately

the same and the histograms has balanced tails, then the data is symmetric.

Examples of skewness and symmetry

These are "perfect" examples of skewness and symmetry. In reality, there may be multiple modes or the mean and median will be similar but not equal. These are provided to give an example.



Measures of variation

An important characteristic of any set of data is the variation in the data. In some data sets, the data values are concentrated closely near the mean; in other data sets, the data values are more widely spread out from the mean. There are five measures of variation: range, standard deviation, variance, interquartile range and coefficient of variation.

The range is the easiest to calculate. It is found by subtracting the maximum value in the data set from the minimum value in the data set. Though the range is easy to calculate, it is very much affected by outliers.

The interquartile range will be discussed in the section on box plots (section 2.3).

The most common measure of variation, or spread, is the standard deviation. The **standard deviation** measures how far data values are from their mean, on average.

Variation within a sample vs. variation between samples

When talking about variable or variability in statistics, there are two different kinds: **variation within a sample** and **variation between samples**. When we discuss finding the standard deviation, range or any measure of variation of a sample, we are discussing variation within a sample. In this case, we are looking at how the data values vary from each other. Most of the time, when we talk about variation this is what we are talking about. We can also talk about how much different samples vary from each other. For example, we could take multiple samples and find the sample mean of each sample. If we talk about how much the means vary from each other, we are discussing variation between samples. We will discuss this specific type of variation in Chapter 6.

The law of large numbers saws that, for random samples, as the sample size increases, then the sample will more closely resemble the population.

For example, as the sample size increases, the sample standard deviation will approach the population standard deviation. Thus, *the variation within the sample will more closely mimic the variation within the population as the sample size increases*. But as the sample size increases, the sample means will approach the population mean. Thus, there will be less variation between the sample means. This means that *the variation between samples decreases, as the sample size increases*. When we discuss **sampling variability**, we are discussing variation between samples.

For this chapter, we are focusing on variation within a sample.

The standard deviation (and variance)

- provides a numerical measure of the overall amount of variation in a data set, and
- can be used to determine whether a particular data value is close to or far from the mean.

The standard deviation provides a measure of the overall variation in a data set

The standard deviation is small when the data are all concentrated close to the mean, exhibiting little variation or spread. The standard deviation is larger when the data values are more spread out from the mean, exhibiting more variation.

Suppose that we are studying the amount of time customers wait in line at the checkout at supermarket A and supermarket B. It is known that the average wait time at both supermarkets is about five minutes. At supermarket A, though, the standard deviation for the wait time is two minutes; at supermarket B the standard deviation for the wait time is four minutes.

Because supermarket B has a higher standard deviation, we know that there is more variation in the wait times at supermarket B. Overall, wait times at supermarket B are more spread out from the average; wait times at supermarket A are more concentrated near the average. This means that at supermarket B, you have a greater chance of having a short wait time, but also a greater chance of having a long wait time, compared to supermarket A. That means the wait times are more volatile at supermarket B. On the other hand, you will be waiting about the same amount of time at supermarket A. That means there are more consistent waits times at supermarket A.

One way, we could summarize the supermarket situation is as follows:

- A typical wait time at supermarket A is 5 minutes give or take 2 minutes. This means that someone typically has to wait 3 to 7 minutes in the checkout line.

- A typical wait time at supermarket B is 5 minutes give or take 4 minutes. This means that someone typically has to wait 1 to 9 minutes in the checkout line.

Here the term “typical” means common, normal. So normally people will wait between 3 to 7 minutes at supermarket A, but there will be some people who only wait 2 minutes and some who wait 10 minutes at the checkout. That is, the typical range only provides a sense of what is going on in the middle of the data, but there are values occurring outside of that range.

For the typical value, you can use any measure of centre. But for the give or take value, you have to use standard deviation. No other measure of variation works.

Calculating the Standard Deviation

The following explains how to calculate the standard deviation by hand. We will be using computer software to do this. Thus it is not important to know this section in detail, but it is helpful to know the basics of how the standard deviation is calculated to help understand what the standard deviation is.

If x is a number, then the difference "x – mean" is called its **deviation**. In a data set, there are as many deviations as there are items in the data set. The deviations are used to calculate the standard deviation. If the numbers belong to a population, in symbols a deviation is x – μ. For sample data, in symbols a deviation is x – x – .

The procedure to calculate the standard deviation depends on whether the numbers are the entire population or are data from a sample. The calculations are similar, but not identical. Therefore the symbol used to represent the standard deviation depends on whether it is calculated from a population or a sample. The lower case letter s represents the sample standard deviation and the Greek letter σ (sigma, lower case) represents the population standard deviation. If the sample has the same characteristics as the population, then s should be a good estimate of σ.

To calculate the standard deviation, we need to calculate the variance first. The **variance** is the **average of the squares of the deviations** (the x – x – values for a sample, or the x – μ values for a population). The symbol σ2 represents the population variance; the population standard deviation *σ* is the square root of the population variance. The symbol *s*2 represents the sample variance; the sample standard deviation *s* is the square root of the sample variance. You can think of the standard deviation as a special average of the deviations.

If the numbers come from a census of the entire **population** and not a sample, when we calculate the average of the squared deviations to find the variance, we divide by *N*, the number of items in the population. If the data are from a **sample** rather than a population, when we calculate the average of the squared deviations, we divide by ***n* – 1**, one less than the number of items in the sample.

Formulas for the Sample Standard Deviation

- $s= \Sigma (x - x -) 2 n-1$
- For the sample standard deviation, the denominator is ***n* - 1**, that is the sample size - 1.

Formulas for the Population Standard Deviation

- $\sigma = \Sigma (x-\mu) 2 N$
- For the population standard deviation, the denominator is *N*, the number of items in the population.

Since the standard deviation is found by square rooting something, the standard deviation is always positive or zero.

Since the variance is the square of the standard deviation, it is not helpful as a descriptive statistic. For example, if you are looking at the weights of basketballs in kg, then the standard deviation will be in kg, while the variance will be in kg^2. Thus the variance is meaningless when trying to interpret the variation in data. It is helpful later on in statistics, but at this point it is not.

In a fifth grade class, the teacher was interested in the average age and the sample standard deviation of the ages of her students. The following data are the ages for a SAMPLE of $n = 20$ fifth grade students. The ages are rounded to the nearest half year:
9; 9.5; 9.5; 10; 10; 10; 10; 10.5; 10.5; 10.5; 10.5; 11; 11; 11; 11; 11; 11; 11.5; 11.5; 11.5;
$x – =$
$9 + 9.5(2) + 10(4) + 10.5(4) + 11(6) + 11.5(3)$
$20 = 10.525$
The average age is 10.53 years, rounded to two places.
The variance may be calculated by using a table. Then the standard deviation is calculated by taking the square root of the variance. We will explain the parts of the table after calculating *s*.

Data	**Freq.**	**Deviations**	***Deviations*2**	**(Freq.) (*Deviations*2)**
x	*f*	(*x* – x –)	$(x – x –)^2$	(*f*)$(x – x –)^2$
9	1	9 – 10.525 = –1.525	$(–1.525)^2$ = 2.325625	1 × 2.325625 = 2.325625
9.5	2	9.5 – 10.525 = –1.025	$(–1.025)^2$ = 1.050625	2 × 1.050625 = 2.101250
10	4	10 – 10.525 = 0.525	$(–0.525)^2$ = 0.275625	4 × 0.275625 = 1.1025
10.5	4	10.5 – 10.525 = 0.025	$(–0.025)^2$ = 0.000625	4 × 0.000625 = 0.0025
11	6	11 – 10.525 = 0.475	$(0.475)^2$ = 0.225625	6 × 0.225625 = 1.35375
11.5	3	11.5 – 10.525 = 0.975	$(0.975)^2$ = 0.950625	3 × 0.950625 = 2.851875
				The total is 9.7375

The sample variance, s^2, is equal to the sum of the last column (9.7375) divided by the total number

of data values minus one (20 – 1):
s 2 = 9.7375 20−1 =0.5125
The **sample standard deviation** *s* is equal to the square root of the sample variance:
s= 0.5125 =0.715891, which is rounded to two decimal places, *s* = 0.72.

Explanation of the standard deviation calculation shown in the table

The deviations show how spread out the data are about the mean. The data value 11.5 is farther from the mean than is the data value 11 which is indicated by the deviations 0.97 and 0.47. A positive deviation occurs when the data value is greater than the mean, whereas a negative deviation occurs when the data value is less than the mean. The deviation is –1.525 for the data value nine. **If you add the deviations, the sum is always zero**. (For [link], there are *n* = 20 deviations.) So you cannot simply add the deviations to get the spread of the data. By squaring the deviations, you make them positive numbers, and the sum will also be positive. The variance, then, is the average squared deviation.

The variance is a squared measure and does not have the same units as the data. Taking the square root solves the problem. The standard deviation measures the spread in the same units as the data.

Notice that instead of dividing by $n = 20$, the calculation divided by $n – 1 = 20 – 1 = 19$ because the data is a sample. For the **sample** variance, we divide by the sample size minus one ($n – 1$). Why not divide by n? The answer has to do with the population variance. **The sample variance is an estimate of the population variance.** Based on the theoretical mathematics that lies behind these calculations, dividing by ($n – 1$) gives a better estimate of the population variance.

The standard deviation, s or σ, is either zero or larger than zero. When the standard deviation is zero, there is no spread; that is, the all the data values are equal to each other. The standard deviation is small when the data are all concentrated close to the mean, and is larger when the data values show more variation from the mean. When the standard deviation is a lot larger than zero, the data values are very spread out about the mean; outliers can make s or σ very large.

Coefficient of variation

The standard deviation is a very good measure of variation, but when comparing two data sets it is not always the best. In particular, if the means of the two data sets are different. Suppose you are comparing the yearly salaries (excluding bonuses) of junior employees versus CEOs at oil and gas companies around Alberta. The yearly salaries for the junior employees will be significantly smaller than the CEOs. Let’s say the average salary for junior employees is $45,000 while for CEOs is $500,000. Now suppose that the standard deviation for both groups is $50,000. If we only looked at the standard deviation, we might say that the variation in both groups is the same. But really variation of $50,000 when the average salary is $45,000 is quite a bit more than for a salary of $500,000. That is, there is more relative variation in the junior employees’ salary. The standard deviation doesn’t capture this difference. But the coefficient of variation does and is a measure of **relative variation**. That is, it takes into account that bigger data values might have a larger standard deviation, but that doesn’t mean it has larger variation.

The coefficient of variation is found by expressing the standard deviation as a percentage of the mean:
Coefficient of Variation = s x – (100%)

In the above example, the coefficient of variation would be:
CofV for Junior employees = 50,000 45,000 (100%) =111.1%
CofV for CEOs = 50,000 5,000,000 (100%) =1%

The larger the coefficient of variation, the larger the relative variation. Thus, as a measure of relative variation, the junior employees have significantly more relative variation (111.11%) compared to the CEOs (1%).

Here are some points about the coefficient of variation:

- The coefficient of variation is not affected by multiplicative changes of scale.
- **The coefficient of variation is used to compare variation between data sets.** This is very important to remember. For multiple data sets, if the means are the same, you can compare the standard deviations. BUT if the means are different, you MUST use the coefficient of variation of compare the variation in the data sets.
- If the standard deviation is larger than the mean, the coefficient of variation is bigger than 100%.

Measure	**When to use**
Range	The range is rarely the best measure of variation to use. But it is a good quick calculation of variation.

Standard deviation	Similar to the mean, this is the most common measure of variation. Also, it is derived from the mean. Therefore, if your best measure of centre is the mean, then the standard deviation is a good complement to it. Further, it is best used when finding the variation for one data set.
Variance	As it the square of the standard deviation, it is NEVER the best measure of variation to use. It is helpful in later topics in statistics though.
Interquartile range	This is a not very well known measure of variation, but it is helpful in describing the range for middle 50% of the data values. Further, it is based on measures of location. Therefore, if your best measure of centre is the median, then the IQR is a good complementary measure of variation.

Coefficient of variation	This is not well known, but it is useful for giving a context free interpretation of variation. It is the best measure to use when comparing the variations of two or more data sets that have different measures of centre.

When to use which measure of variation

Suppose you are looking at two companies and each company has 24 employees. At one company, everybody except the CEO makes $30,000. The CEO makes $490,000. Thus, the data values would be

$30,000; $30,000; $30,000; $30,000; $30,000; … ; $490,000

The second company has an interesting policy. Everybody who starts at the company makes $30,000 a year, but as soon as someone else gets hired, they get paid $20,000 more. They only hire one person at a time. So, the first person who was hired started at $30,000, then when a second person got hired, the first person's salary was raised to $50,000. When a third person got hired, the first person's salary was raised to $70,000 while the salary of the second person hired was raised to $50,000. This has been done 23 times. Therefore, their data values (i.e. salaries) would look like this:

$30,000 $50,000; $70,000; $90,000; $110,000; … ;$490,000

Without doing any calculations, we can see that company one has fairly consistent salaries except for the CEO. While company two has salaries that are more spread out.

The following table provides the count (i.e. sample size), mean, and the measures of variation for the two companies.

	Company One	Company Two
Count	24	24
Mean	49,166.67	260,000.00
Range	460,000	460,000
Population standard deviation	91,820.10	138,443.73
Coefficient of Variation	190.98%	54.39%

In the table above, notice that the range is the same for the two data sets. If we only looked at the range, this would give a false sense that the amount of variation in the two data sets is the same, but we know it isn't.

The standard deviation is measuring how much, on average, the data values vary from the mean. For company one, 23 of the 24 data values deviate the same amount from the mean ($49,166.67 – $30,000 = $19,166.67) with only the $490,000 deviating a large amount from the mean.

For company two, two data values only deviate by only $10,000 ($250,000 and $270,000) while two data values deviate by a whopping $230,000 ($30,000 and $490,000).

In company one, 23 out of 24 data values deviate by less than $20,000. But for company two, only 2 out of 24 deviate by less than $20,000. This suggests that company one will have a smaller standard deviation than company two because there is less average deviation. This is supported by MegaStat, which shows that the population standard deviation for company one is $91,920.10 versus company two, which has a population standard deviation of $138,443.73.

Notice that even though company one has an outlier (the CEO's salary), the standard deviation is less than company two. That is, the average variation from the mean is less for company one. Thus, *the presence of an outlier does not necessarily result in a larger standard deviation*.

The story is different when we look at the coefficient of variation. For company one, it is

190.98%. While for company two, it is 54.39%. This means that company one has larger relative variation than company two. This is because company two has a higher mean than company one and thus the variation, relative to the mean, isn’t as large as it is in company one.

In this situation, the best measure of variation to use would be the coefficient of variation as we are comparing two data sets with two different means. Based on this, company one has larger relative variation than company two.

Notice that variance is not discussed here. As stated above, the variance is the square of the standard deviation. Therefore, the units for variance in this example would be $ˆ2, which makes no sense. Again, variance is not a useful descriptive statistic.

Common Mistake

Variation and variance might seem like the same word but they aren’t. **Variation** is a general term used to discuss how much the data values vary from each other, how much spread there is in the data, how consistent the data is, how volatile or risky the data is, and how much deviation there is in the data values. It is an umbrella term. **Variance** is a specific type of variation. It specifically refers to the square of the standard deviation. Therefore, it is *incorrect* to say, “There is a lot of variance in the data” or “The best measure of variance is …”.

Optional section: Comparing Values from Different Data Sets

The standard deviation is useful when comparing data values that come from different data sets. If the data sets have different means and standard deviations, then comparing the data values directly can be misleading.

- For each data value, calculate how many standard deviations away from its mean the value is.
- Use the formula: value = mean + (#ofSTDEVs)(standard deviation); solve for #ofSTDEVs.
- #ofSTDEVs = value – mean standard deviation
- Compare the results of this calculation.

#ofSTDEVs is often called a "*z*-score"; we can use the symbol *z*. In symbols, the formulas become:

Sample	$x = \bar{x} + zs$	$z = x - x \quad s$
Population	$x = \mu + z\sigma$	$z = x - \mu\ \sigma$

Two students, John and Ali, from different high schools, wanted to find out who had the highest GPA when compared to his school. Which student had the highest GPA when compared to his school?

Student	GPA	School Mean GPA	School Standard Deviation
John	2.85	3.0	0.7
Ali	77	80	10

For each student, determine how many standard deviations (#ofSTDEVs) his GPA is away from the average, for his school. Pay careful attention to signs when comparing and interpreting the answer.

z= # of STDEVs= value –mean
standard deviation = x-μ σ

For John, z=#ofSTDEVs= 2.85–3.0 0.7 =–
0.21

For Ali, $z = \#ofSTDEVs = \ 77 - 80\ 10\ = $-0.3

John has the better GPA when compared to his school because his GPA is 0.21 standard deviations **below** his school's mean while Ali's GPA is 0.3 standard deviations **below** his school's mean.

John's *z*-score of –0.21 is higher than Ali's *z*-score of –0.3. For GPA, higher values are better, so we conclude that John has the better GPA when compared to his school.

Try It

Two swimmers, Angie and Beth, from different teams, wanted to find out who had the fastest time for the 50 meter freestyle when compared to her team. Which swimmer had the fastest time when compared to her team?

Swimmer	**Time (seconds)**	**Team Mean Time**	**Team Standard Deviation**

Angie	26.2	27.2	0.8
Beth	27.3	30.1	1.4

For Angie: $z = 26.2 - 27.2\ 0.8 = -1.25$

For Beth: $z = 27.3 - 30.1\ 1.4 = -2$

Distributions

Now that we have learned about determining shape (histogram), centre (mean, median or mode), and variation (standard deviation, coefficient of variation and range), we can now describe the distribution of a data set.

In [link], we examined the salaries for two different companies.

Though we have not done the histogram for either of these data sets, we can imagine what they will look like to determine the shape. Company A will have one peak at $30,000 with an outlier at $490,000. This will make it skewed to the right. For Company B each data value has the same frequency, which makes the data uniform.

For company A, we would describe the distribution of salaries to be skewed to the right(shape), centred at $49,166.67 (mean) and have variation of $91,820.10 (standard deviation).

For company B, we would describe the distribution of salaries to be uniform(shape), centred at $260,000 (mean) and have variation of $138,443.73 (standard deviation).

References

Data from The World Bank, available online at http://www.worldbank.org (accessed April 3, 2013).

"Demographics: Obesity – adult prevalence rate." Indexmundi. Available online at http://www.indexmundi.com/g/r.aspx?t=50&v=2228&l=en (accessed April 3, 2013).

Chapter Review

The mean and the median can be calculated to help you find the "center" of a data set. The mean is the best estimate for the actual data set, but the median is the best measurement when a data set contains several outliers or extreme values. The mode will tell you the most frequently occuring datum (or data) in your data set. The mean, median, and mode are extremely helpful when you need to analyze your data, but if your data set consists of ranges which lack specific values, the mean may seem impossible to calculate. However, the mean can be approximated if you add the lower boundary with the upper boundary and divide by two to find the midpoint of each interval. Multiply each midpoint by the number of values found in the corresponding range. Divide the sum of these values by the total number of data values in the set.

The standard deviation can help you calculate the spread of data. There are different equations to use if are calculating the standard deviation of a sample or of a population.

- The Standard Deviation allows us to compare individual data or classes to the data set mean numerically.
- s = Σ (x− x –) 2 n−1 or s = Σ f (x− x –) 2 n−1 is the formula for calculating the standard deviation of a sample. To calculate the standard deviation of a population, we would use the population mean, μ, and the formula σ = Σ (x−μ) 2 N or σ = Σ f (x−μ) 2 N .

Use the following information to answer the next three

exercises: The following data show the lengths of boats moored in a marina. The data are ordered from smallest to largest: 1617192020212324252525262627272728293032333333435373940

Calculate the mean.

Mean: 16 + 17 + 19 + 20 + 20 + 21 + 23 + 24 + 25 + 25 + 25 + 26 + 26 + 27 + 27 + 27 + 28 + 29 + 30 + 32 + 33 + 33 + 34 + 35 + 37 + 39 + 40 = 738;

738 27 = 27.33

Identify the median.

Median = 27

Identify the mode.

The most frequent lengths are 25 and 27, which occur three times. Mode = 25, 27

Use the following information to answer the next three

exercises: Sixty-five randomly selected car salespersons were asked the number of cars they generally sell in one week. Fourteen people answered that they generally sell three cars; nineteen generally sell four cars; twelve generally sell five cars; nine generally sell six cars; eleven generally sell seven cars. Calculate the following:

sample mean = $\bar{x}$ = _____

Mean = (14*3+19*4+12*5+9*6+11*7)/65 = 4.75

median = _____

4

mode = _____

Mode = 4 (occurs 19 times)

The following data are the distances between 20 retail stores and a large distribution center. The distances are in miles.
29; 37; 38; 40; 58; 67; 68; 69; 76; 86; 87; 95; 96; 96; 99; 106; 112; 127; 145; 150

Use a computer to find the standard deviation and round to the nearest tenth.

$s = 34.5$

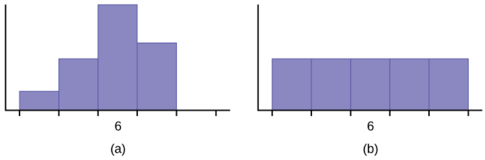
Bringing It Together

Javier and Ercilia are supervisors at a shopping mall. Each was given the task of estimating the mean distance that shoppers live from the mall. They each randomly surveyed 100 shoppers. The samples yielded the following information.

	Javier	Ercilia
$\bar{x}$	6.0 km	6.0 km
s	4.0 km	7.0 km

1. How can you determine which survey was correct ?
2. Explain what the difference in the results of the surveys implies about the data.
3. If the two histograms depict the

distribution of values for each supervisor, which one depicts Ercilia's sample? How do you know?



1. It is difficult to determine which survey is correct. Both surveys include the same number of shoppers and the shoppers were randomly selected. We could look at how the random selection was done to see if one of the sampling techniques would result in a more representative sample. But if they used the same sampling technique, there is no way to know which sample is right. The only way would be to take another, larger sample and see which of the two supervisor's samples most closely matches that sample. But really we expect there to be sampling variability so it is not really an appropriate question to ask which is "correct".
2. Since the mean is the same for both samples, this suggests that it is fair to say that on average shoppers travel 6.0 km to the mall. But the standard deviations are different. This suggests that it is not yet clear how much variation there is from the

6.0km.

3. Ercilia's data has a larger standard deviation. Therefore, on average, the data needs to be more spread out from the mean than Javier's. This suggests (b) is the answer.

Use the following information to answer the next three exercises: We are interested in the number of years students in a particular elementary statistics class have lived in California. The information in the following table is from the entire section.

Number of years	**Frequency**	**Number of years**	**Frequency**
			Total = 20
7	1	22	1
14	3	23	1
15	1	26	1
18	1	40	2
19	4	42	2
20	3		

What is the mode?

1. 19
2. 19.5
3. 14 and 20
4. 22.65

Mode = 19 (occurs 4 times)

Is this a sample or the entire population?

1. sample
2. entire population
3. neither

---

b

A survey of enrollment at 35 community colleges across the United States yielded the following figures:

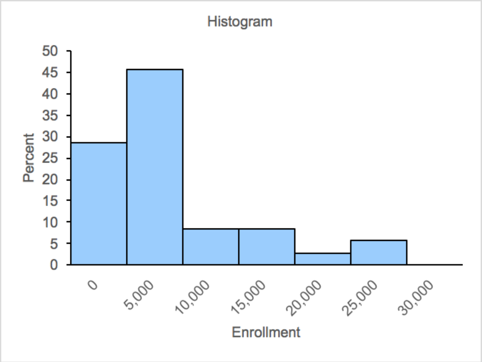
6414; 1550; 2109; 9350; 21828; 4300; 5944; 5722; 2825; 2044; 5481; 5200; 5853; 2750; 10012; 6357; 27000; 9414; 7681; 3200; 17500; 9200; 7380; 18314; 6557; 13713; 17768; 7493; 2771; 2861; 1263; 7285; 28165; 5080; 11622

1. Organize the data into a chart with six

intervals of equal width. Label the two columns "Enrollment" and "Frequency."
2. Construct a histogram of the data.
3. What is the shape of the data? What does the shape tell you about the enrollment at these community colleges?
4. What is the best measure of centre for this data and why? State the measure.
5. What is the best measure of variation for this data and why? State the measure.
6. If you were to build a new community college, what is the typical range for the enrollment? Why would this information be helpful? What caveats would you want to think about when you look at this typical range?

1. Enrollment	Frequency	
0-4999	10	
5000-9999	16	
10000-14999	3	
15000-19999	3	
20000-24999	1	
25000-29999	2	

2. **Histogram for enrollment at community colleges.**



3. The shape is skewed to the right which means that there a few community colleges that have greater enrollment compared to most of the other colleges in the sample.
4. Since the mean (8628.74) is being skewed (as it is larger than the median of 6,414), the median is the best measure of centre.
5. Since we are only looking at one data set, the standard deviation is a good measure of variation. It is 6,943.88.
6. The typical range is 6,414 +/- 6,943.88 = -529.88 to 13,357.88. As there can't be negative students enrolled, the typical range is 0 students to 13,357.88. Though there could be multiple caveats, one concern is the rather large variation in the data. This means that community colleges have very different enrollment rates. Perhaps looking at community colleges that are similar to the one I would like to

open would be more beneficial as that population would be more representative of my community college.

You work for a soda pop company that is producing a new label for their Asian market. Three different labels your company is considering are the same, except the colours are different. The colour choices are blue, green and orange.

To determine which label consumers prefer, focus groups were done. One such focus group asked 15 participants to rate the cans from 1 to 10. A score of 1 means they hated the label and 10 means they loved the label. The results follow.

Participant	Blue Label	Green Label	Orange Label
1	1	10	6
2	4	8	7
3	2	9	7
4	6	3	8
5	1	8	6

6	1		7		7	
7	1		3		7	
8	4		9		8	
9	1		10		9	
10	7		4		6	
11	4		7		6	
12	5		6		7	
13	6		9		8	
14	4		4		6	
15	6		8		7	

Which label would you recommend as the new label for the Asian market? Support your decision using the data.

Label 1 is excluded as most people don’t like it. The mean for label 2 and label 3 is the same. Label 2 could be considered the better label because more people love it than label 3, but more people hate it. Label 3 could be considered a better label because the variation is less - nobody hates it, but nobody loves it. (Note: Even though you are comparing two data sets, it is ok to look only at the standard deviation instead of the coefficient of variation in this situation. Why?).

Choosing label 2 has greater risk (love/hate relationship). Choosing label 3 has less risk (most people like it).

Three publicly traded telecommunications companies reported their monthly profit for the last year. The results are presented below.

	Company A	Company B	Company C
Mean	$10,930	$13,000	$34,450
Median	$9,390	$13,500	$34,450
Mode	None	$13,000 and $20,000	$33,880
Standard deviation	$4,196	$9,360	$4,116
Range	$15,050	$42,150	$16,400

1. Donna is close to retirement and wants to invest in one of the three companies. She doesn't want to see her investment drop significantly as she doesn't want to see her retirement savings dwindle. Which company would you recommend she invest in and why?
2. What information is missing from the list that you might want to have to help you answer the above question?
3. What information below is not necessary for making this decision?

Note that this question is about risk, i.e. variation.

1. Any answer requires that you examine the amount of variation in the data set. The coefficient of variation is the best measure to use to compare the variation as the means are different.

	Company A	Company B	Company C
Coefficient of variation	38.39%	72%	11.95%

2. The information provided is only for one year. It would be helpful to know about their changes over more than one year. Quartiles aren't provided. They could help examine the variation as well.
3. The median and the mode are not relevant as this is a question about variation. The mean is only required as it is needed to find the coefficient of variation.

Glossary

Frequency Table
: a data representation in which grouped data is displayed along with the corresponding

frequencies

Mean

a number that measures the central tendency of the data; a common name for mean is 'average.' The term 'mean' is a shortened form of 'arithmetic mean.' By definition, the mean for a sample (denoted by x¯) is x – =
Sum of all values in the sample
Number of values in the sample , and the mean for a population (denoted by μ) is μ =
Sum of all values in the population
Number of values in the population .

Median

a number that separates ordered data into halves; half the values are the same number or smaller than the median and half the values are the same number or larger than the median. The median may or may not be part of the data.

Midpoint

the mean of an interval in a frequency table

Mode

the value that appears most frequently in a set of data

Visual representations of categorical data

Below are tables comparing the number of part-time and full-time students at De Anza College and Foothill College enrolled for the spring 2010 quarter. The tables display counts (frequencies) and percentages or proportions (relative frequencies). The percent columns make comparing the same categories in the colleges easier. Displaying percentages along with the numbers is often helpful, but it is particularly important when comparing sets of data that do not have the same totals, such as the total enrollments for both colleges in this example. Notice how much larger the percentage for part-time students at Foothill College is compared to De Anza College.

De Anza College		
	Number	Percent
Full-	9,200	40.9%

Foothill College		
	Number	Percent
Full-	4,059	28.6%

time		
Part-time	13,296	59.1%
Total	22,496	100%

time		
Part-time	10,124	71.4%
Total	14,183	100%

Fall Term 2007 (Census day)

Tables are a good way of organizing and displaying data. But graphs can be even more helpful in understanding the data. There are no strict rules concerning which graphs to use. Two graphs that are used to display categorical data are pie charts and bar graphs.

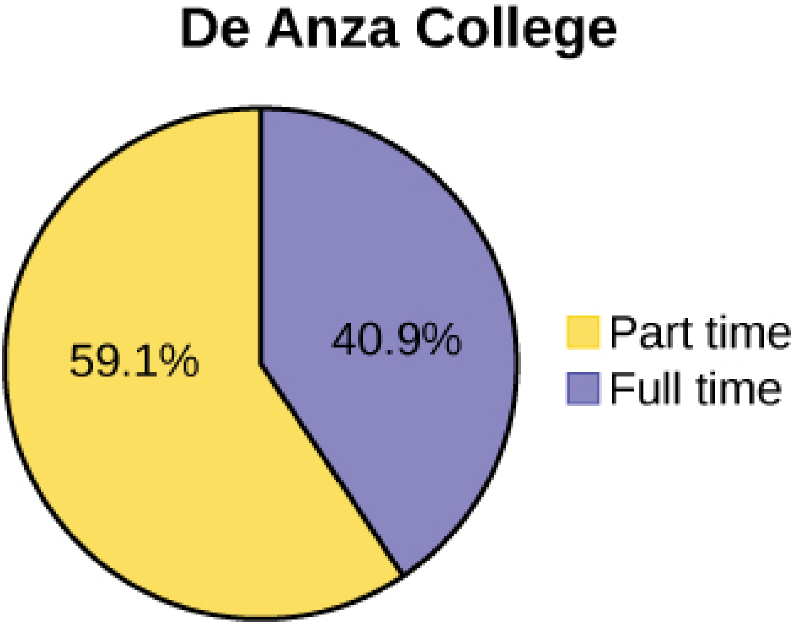
In a **pie chart**, categories of data are represented by wedges in a circle and are proportional in size to the percent of individuals in each category.

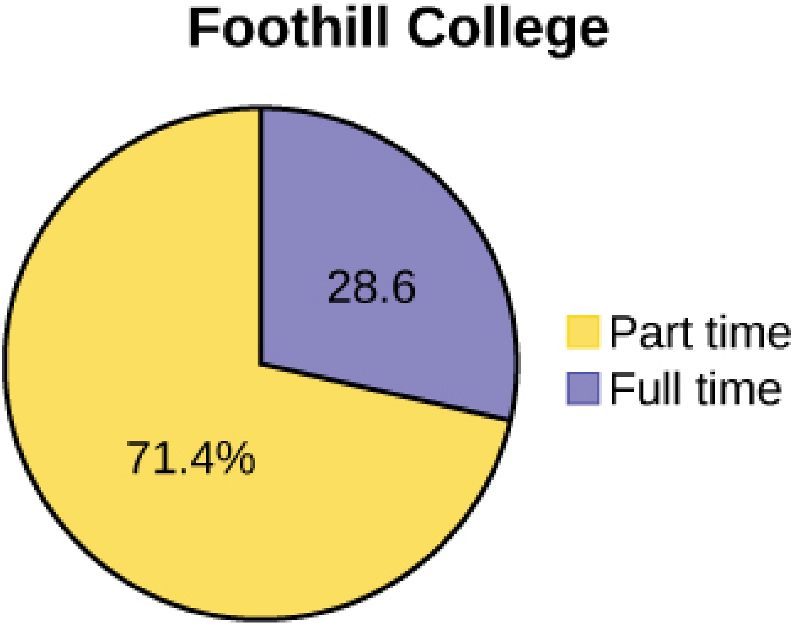
In a **bar graph**, the length of the bar for each category is proportional to the number or percent of individuals in each category. Bars may be vertical or horizontal.

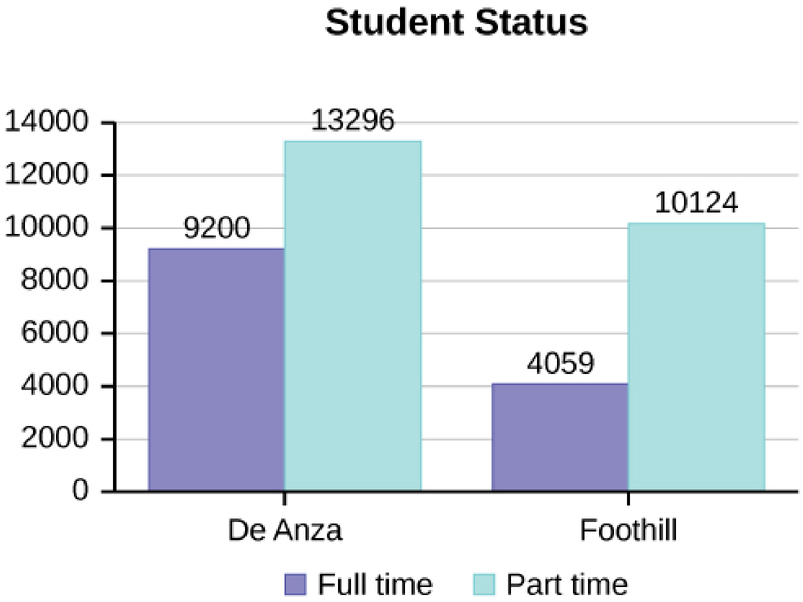
Look at [link] and [link] and determine which graph (pie or bar) you think displays the comparisons better.

It is a good idea to look at a variety of graphs to see which is the most helpful in displaying the data. We might make different choices of what we think is the “best” graph depending on the data and the context.

Our choice also depends on what we are using the data for.







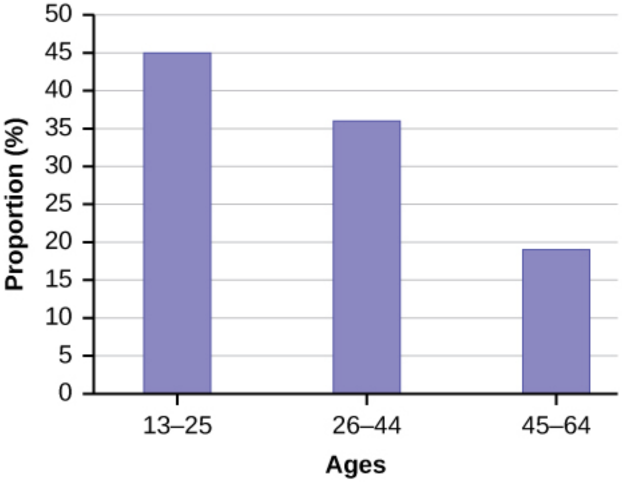
Visual Representations of Quantitative Data

Bar Graphs

Bar graphs can also be used to summarize discrete quantitative data and categorical data. Bar graphs consist of bars that are separated from each other. The bars can be rectangles or they can be rectangular boxes (used in three-dimensional plots), and they can be vertical or horizontal. The **bar graph** shown in [link] has age groups represented on the ***x*-axis** and proportions on the ***y*-axis**.

By the end of 2011, Facebook had over 146 million users in the United States. [link] shows three age groups, the number of users in each age group, and the proportion (%) of users in each age group. Construct a bar graph using this data.

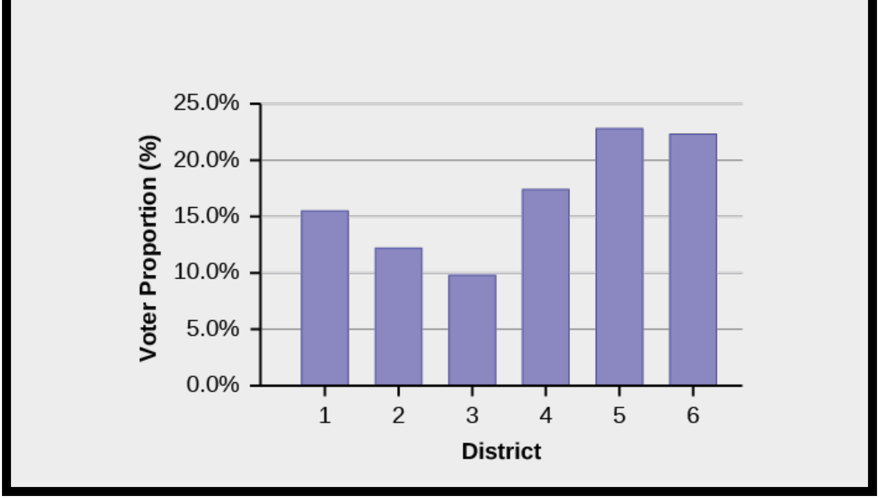
Age groups	Number of Facebook users	Proportion (%) of Facebook users
13–25	65,082,280	45%
26–44	53,300,200	36%
45–64	27,885,100	19%



Try it

Park city is broken down into six voting districts. The table shows the percent of the total registered voter population that lives in each district as well as the percent total of the entire population that lives in each district. Construct a bar graph that shows the registered voter population by district.

District	Registered voter population	Overall city population
1	15.5%	19.4%
2	12.2%	15.6%
3	9.8%	9.0%
4	17.4%	18.5%
5	22.8%	20.7%
6	22.3%	16.8%



Frequency tables

Twenty students were asked how many hours they worked per day. Their responses, in hours, are as follows: 56332475235654435253.

[link] lists the different data values in ascending order and their frequencies.

DATA VALUE	FREQUENCY
2	3
3	5
4	3
5	6
6	2

7	1	

Frequency Table of Student Work Hours

A **frequency** is the number of times a value of the data occurs. According to [link], there are three students who work two hours, five students who work three hours, and so on. The sum of the values in the frequency column, 20, represents the total number of students included in the sample.

A **relative frequency** is the ratio (fraction or proportion) of the number of times a value of the data occurs in the set of all outcomes to the total number of outcomes. To find the relative frequencies, divide each frequency by the total number of students in the sample–in this case, 20. Relative frequencies can be written as fractions, percents, or decimals.

DATA VALUE	FREQUENCY	RELATIVE FREQUENCY
2	3	3 20 or 0.15
3	5	5 20 or 0.25
4	3	3 20 or 0.15
5	6	6 20 or 0.30

6	2	2 20 or 0.10
7	1	1 20 or 0.05

Frequency Table of Student Work Hours with Relative Frequencies

The sum of the values in the relative frequency column of [link] is 20 20 , or 1.

Cumulative relative frequency is the accumulation of the previous relative frequencies. To find the cumulative relative frequencies, add all the previous relative frequencies to the relative frequency for the current row, as shown in [link].

DATA VALUE	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
2	3	3 20 or 0.15	0.15
3	5	5 20 or 0.25	0.15 + 0.25 = 0.40
4	3	3 20 or 0.15	0.40 + 0.15 = 0.55
5	6	6 20 or 0.30	0.55 + 0.30 = 0.85
6	2	2 20 or 0.10	0.85 + 0.10

= 0.95

7	1	1 20 or 0.05	0.95 + 0.05 = 1.00

Frequency Table of Student Work Hours with Relative and Cumulative Relative Frequencies

The last entry of the cumulative relative frequency column is one, indicating that one hundred percent of the data has been accumulated.

NOTE
Because of rounding, the relative frequency column may not always sum to one, and the last entry in the cumulative relative frequency column may not be one. However, they each should be close to one.

[link] represents the heights, in inches, of a sample of 100 male semiprofessional soccer players.

HEIGHTS (INCHES)	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE

			FREQUENCY
60–61.99	5	5 100 = 0.05	0.05
62–63.99	3	3 100 = 0.03	0.05 + 0.03 = 0.08
64–65.99	15	15 100 = 0.15	0.08 + 0.15 = 0.23
66-67.99	40	40 100 = 0.40	0.23 + 0.40 = 0.63
68–69.99	17	17 100 = 0.17	0.63 + 0.17 = 0.80
70–71.99	12	12 100 = 0.12	0.80 + 0.12 = 0.92
72–73.99	7	7 100 = 0.07	0.92 + 0.07 = 0.99
74–75.99	1	1 100 = 0.01	0.99 + 0.01 = 1.00
	Total = 100	**Total = 1.00**	

Frequency Table of Soccer Player Height

The data in this table have been **grouped** into the following intervals:

- 60 to 61.99 inches
- 62 to 63.99 inches
- 64 to 65.99 inches
- 66 to 67.99 inches
- 68 to 69.99 inches
- 70 to 71.99 inches

- 72 to 73.99 inches
- 74 to 75.99 inches

In this sample, there are **five** players whose heights fall within the interval 59.95–61.95 inches, **three** players whose heights fall within the interval 61.95–63.95 inches, **15** players whose heights fall within the interval 63.95–65.95 inches, **40** players whose heights fall within the interval 65.95–67.95 inches, **17** players whose heights fall within the interval 67.95–69.95 inches, **12** players whose heights fall within the interval 69.95–71.95, **seven** players whose heights fall within the interval 71.95–73.95, and **one** player whose heights fall within the interval 73.95–75.95. All heights fall between the endpoints of an interval and not at the endpoints.

From [link], find the percentage of heights that are less than 65.95 inches.

If you look at the first, second, and third rows, the heights are all less than 65.95 inches. There are $5 + 3 + 15 = 23$ players whose heights are less than 65.95 inches. The percentage of heights less than 65.95 inches is then 23 100 or 23%. This percentage is the cumulative relative frequency entry in the third row.

Try It

[link] shows the amount, in inches, of annual rainfall in a sample of towns.

Rainfall (Inches)	Frequency	Relative Frequency	Cumulative Relative Frequency
3–4.99	6	6 50 = 0.12	0.12
5–6.99	7	7 50 = 0.14	0.12 + 0.14 = 0.26
7–9.99	15	15 50 = 0.30	0.26 + 0.30 = 0.56
10–11.99	8	8 50 = 0.16	0.56 + 0.16 = 0.72
12–12.99	9	9 50 = 0.18	0.72 + 0.18 = 0.90

13–14.99	5		5 50 = 0.10	0.90 + 0.10 = 1.00
	Total = 50		Total = 1.00	

From [link], find the percentage of rainfall that is less than 9.99 inches.

Try It Solutions

0.56 or 56%

From [link], find the percentage of heights that fall between 61.95 and 65.95 inches.

Add the relative frequencies in the second and third rows: 0.03 + 0.15 = 0.18 or 18%.

Try It

From [link], find the percentage of rainfall that is between 7.00 and 12.99 inches.

Try It Solutions

$0.30 + 0.16 + 0.18 = 0.64$ or 64%

Use the heights of the 100 male semiprofessional soccer players in [link]. Fill in the blanks and check your answers.

1. The percentage of heights that are from 67.95 to 71.95 inches is: ___.
2. The percentage of heights that are from 67.95 to 73.95 inches is: ___.
3. The percentage of heights that are more than 65.95 inches is: ___.
4. The number of players in the sample who are between 61.95 and 71.95 inches tall is: ___.
5. What kind of data are the heights?
6. Describe how you could gather this data (the heights) so that the data are characteristic of all male semiprofessional soccer players.

Remember, you **count frequencies**. To find the relative frequency, divide the frequency by the total number of data values. To find the cumulative relative frequency, add all of the previous relative frequencies to the relative frequency for the current row.

1. 29%
2. 36%
3. 77%
4. 87
5. quantitative continuous
6. get rosters from each team and choose a simple random sample from each

Nineteen people were asked how many miles, to the nearest mile, they commute to work each day. The data are as follows: 2 5 7 3 2 10 18 15 20 7 10 18 5 12 13 12 4 5 10. [link] was produced:

DATA	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
3	3	3 19	0.1579
4	1	1 19	0.2105
5	3	3 19	0.1579

7	2	2 19	0.2632
10	3	4 19	0.4737
12	2	2 19	0.7895
13	1	1 19	0.8421
15	1	1 19	0.8948
18	1	1 19	0.9474
20	1	1 19	1.0000

Frequency of Commuting Distances

1. Is the table correct? If it is not correct, what is wrong?
2. True or False: Three percent of the people surveyed commute three miles. If the statement is not correct, what should it be? If the table is incorrect, make the corrections.
3. What fraction of the people surveyed commute five or seven miles?
4. What fraction of the people surveyed commute 12 miles or more? Less than 12 miles? Between five and 13 miles (not including five and 13 miles)?

1. No. The frequency column sums to 18, not 19. Not all cumulative relative frequencies are correct.
2. False. The frequency for three miles should be one; for two miles (left out), two. The cumulative relative frequency

column should read: 0.1052, 0.1579, 0.2105, 0.3684, 0.4737, 0.6316, 0.7368, 0.7895, 0.8421, 0.9474, 1.0000.

3. 519
4. 719, 1219, 719

Try It

[link] represents the amount, in inches, of annual rainfall in a sample of towns. What fraction of towns surveyed get at least 12 inches of rainfall each year?

Try It Solutions

14 50

Histograms

In the introduction, the idea of distribution was introduced. The distribution refers to the shape, centre and variation of quantitative data. To determine the shape of the data, we need to look at a visual representation of the data. The best visual representation to look at is the histogram.

Bar graphs and histograms look very similar. They both have bars whose heights represent the frequency of the data. But bar graphs are used for categorical data and discrete quantitative data (i.e. whole number data). Histograms are used for continuous quantitative data (i.e. numbers with decimals) and sometimes discrete quantitative data as well. Since there is a gap between categories and whole numbers, the bars in bar graphs do not touch. But for continuous data, there is no gap between the numbers, so the bars for histograms do touch.

For most of the work you do in this book, you will use a histogram to display the data. One advantage of a histogram is that it can readily display large data sets. The following explains how to make a histogram by hand, but you can use statistical software to do this quite quickly.

A **histogram** consists of contiguous (adjoining) boxes. It has both a horizontal axis and a vertical axis. The horizontal axis is labeled with what the data represents (for instance, distance from your home to school). The vertical axis is labeled either **frequency** or **relative frequency** (or percent frequency or probability). The graph will have the same shape with either label. The histogram (like the stemplot) can give you the shape of the data, the center, and the spread of the data.

The relative frequency is equal to the frequency for an observed value of the data divided by the total number of data values in the sample.(Remember, frequency is defined as the number of times an answer occurs.) If:

- f = frequency
- n = total number of data values (or the sum of the individual frequencies), and
- RF = relative frequency,

then:
RF = f n

For example, if three students in Mr. Ahab's English class of 40 students received from 90% to 100%, then, $f = 3$, $n = 40$, and RF = fn = 340 = 0.075. 7.5% of the students received 90–100%. 90–100% are quantitative measures.

To construct a histogram, first decide how many **bars** or **intervals**, also called classes, represent the data. Many histograms consist of five to 15 bars or classes for clarity. The number of bars needs to be chosen. Choose a starting point for the first interval to be less than the smallest data value. A **convenient starting point** is a lower value carried out to one more decimal place than the value with the most decimal places. For example, if the value with the most decimal places is 6.1 and this is the smallest value, a convenient starting point is 6.05 (6.1 – 0.05 = 6.05). We say that 6.05 has more precision. If the value with the most decimal places is 2.23 and the lowest value is 1.5, a convenient starting point is 1.495 (1.5 – 0.005 = 1.495). If the value with the most decimal places is 3.234 and the lowest value is 1.0, a convenient starting point is 0.9995 (1.0 – 0.0005 = 0.9995). If all the data happen to be integers and the smallest value is two, then a convenient starting point is 1.5 (2 – 0.5 = 1.5). Also, when the starting point and other boundaries are carried to one additional decimal place, no data value will fall on a boundary. The next two examples go into detail about how to construct a histogram using continuous data and how to create a histogram using discrete data.

The following data are the heights (in inches to the nearest half inch) of 100 male semiprofessional soccer players. The heights are **continuous** data, since height is measured.

60; 60.5; 61; 61; 61.5

63.5; 63.5; 63.5

64; 64; 64; 64; 64; 64; 64; 64.5; 64.5; 64.5; 64.5; 64.5; 64.5; 64.5; 64.5

66; 66; 66; 66; 66; 66; 66; 66; 66; 66; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5; 66.5;

67; 67; 67; 67; 67; 67; 67; 67; 67; 67; 67; 67; 67.5; 67.5; 67.5; 67.5; 67.5; 67.5; 67.5

68; 68; 69; 69; 69; 69; 69; 69; 69; 69; 69; 69; 69.5; 69.5; 69.5; 69.5; 69.5
70; 70; 70; 70; 70; 70; 70.5; 70.5; 70.5; 71; 71; 71
72; 72; 72; 72.5; 72.5; 73; 73.5
74

The smallest data value is 60. Since the data with the most decimal places has one decimal (for instance, 61.5), we want our starting point to have two decimal places. Since the numbers 0.5, 0.05, 0.005, etc. are convenient numbers, use 0.05 and subtract it from 60, the smallest value, for the convenient starting point.

$60 – 0.05 = 59.95$ which is more precise than, say, 61.5 by one decimal place. The starting point is, then, 59.95.

The largest value is 74, so $74 + 0.05 = 74.05$ is the ending value.

Next, calculate the width of each bar or class interval. To calculate this width, subtract the starting point from the ending value and divide by the number of bars (you must choose the number of bars you desire). Suppose you choose eight bars.

$$74.05-59.958=1.76$$

NOTE

We will round up to two and make each bar or class interval two units wide. Rounding up to two is one way to prevent a value from falling on a boundary. Rounding to the next number is often necessary even if it goes against the standard rules of rounding. For this example, using 1.76 as the width would also work. A guideline that is followed by some for the width of a bar or class interval is to take the square root of the number of data values and then round to the nearest whole number, if necessary. For example, if there are 150 values of data, take the square root of 150 and round to 12 bars or intervals.

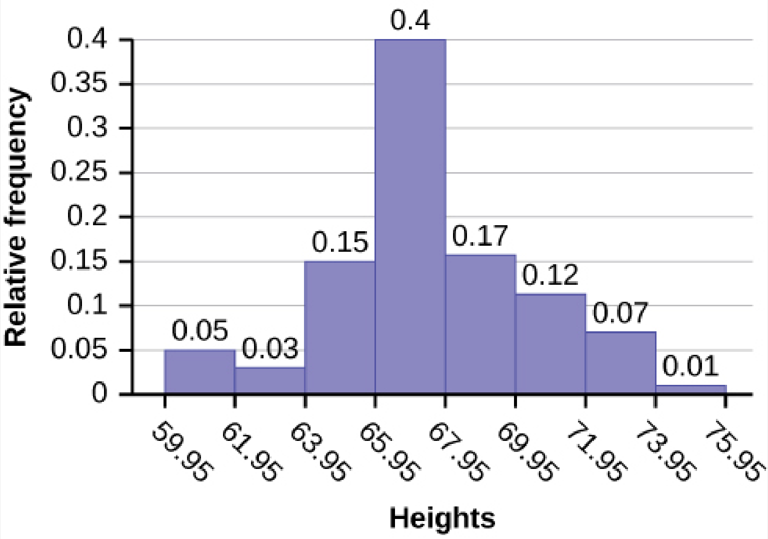
The boundaries are:

- 59.95
- 59.95 + 2 = 61.95
- 61.95 + 2 = 63.95
- 63.95 + 2 = 65.95
- 65.95 + 2 = 67.95
- 67.95 + 2 = 69.95
- 69.95 + 2 = 71.95
- 71.95 + 2 = 73.95
- 73.95 + 2 = 75.95

The heights 60 through 61.5 inches are in the interval 59.95–61.95. The heights that are 63.5 are in the interval 61.95–63.95. The heights that are 64 through 64.5 are in the interval 63.95–65.95. The heights 66 through 67.5 are in the interval 65.95–67.95. The heights 68 through 69.5 are in the interval 67.95–69.95. The heights 70 through

71 are in the interval 69.95–71.95. The heights 72 through 73.5 are in the interval 71.95–73.95. The height 74 is in the interval 73.95–75.95.

The following histogram displays the heights on the *x*-axis and relative frequency on the *y*-axis.



Titles, labelling and numbering of visual representations

Visual representations should be numbered. As they are images, they would be numbered as figures. For example, a histogram would be numbered “Figure 3”. This means it is the third image in the document. This makes it easier to refer back to: “In Figure 3, we can see that …”

The title of the visual representation includes the name of the visual representation and the context:

"Histogram of …".

The label that goes along the axis includes the variable and the unit: Variable (unit).

These three aspects combined will make it easy to refer to the image and let the reader of the image know what the image is about.

A frequency table would be similarly titled and labelled, but since it is a table and not an image, it would be referred to as "Table 4" (meaning the fourth table in the document).

As you look through this textbook, notice how all of the images and tables are numbered as described above.

Try It

The following data are the shoe sizes of 50 male students. The sizes are continuous data since shoe size is measured. Construct a histogram and calculate the width of each bar or class interval. Suppose you choose six bars.

9; 9; 9.5; 9.5; 10; 10; 10; 10; 10; 10; 10.5;
10.5; 10.5; 10.5; 10.5; 10.5; 10.5; 10.5
11; 11; 11; 11; 11; 11; 11; 11; 11; 11; 11; 11;
11; 11.5; 11.5; 11.5; 11.5; 11.5; 11.5; 11.5
12; 12; 12; 12; 12; 12; 12; 12.5; 12.5; 12.5;
12.5; 14

Smallest value: 9

Largest value: 14

Convenient starting value: 9 – 0.05 = 8.95

Convenient ending value: 14 + 0.05 = 14.05

$14.05 - 8.95\ 6 = 0.85$

The calculations suggests using 0.85 as the width of each bar or class interval. You can also use an interval with a width equal to one.

Shape

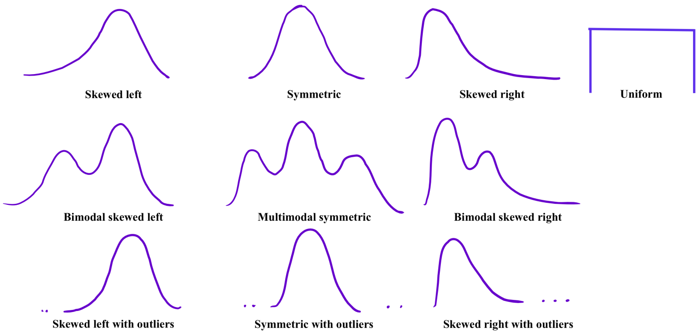
The shape of the data helps us understand what kind of pattern the data has. For example, if all of the data values have the same frequency, then the shape will be distinct (it is called uniform). If the data has a skew in it, then that helps us understand the measure of centre better (to be discussed in the next section). Overall, the shape helps us see how the data is behaving. Data that has similar shapes will behave in similar ways.

The shape of the data set is determined by looking at a visual representation of the data and usually the histogram. Common ways of describing the shape include whether it is symmetrical or not, how many distinct peaks it has (unimodal, bimodal, multimodal), and whether the data has a tail only on one side (skew).

- Data is symmetric if the shape is same on both sides of centre.
- Skewed data has a "tail" on one side. This means that there are some data values that are far from the centre but only one one side. This is a type of non-symmetric data.
- For a histogram, the term "modal" refers to the number of distinct peaks. You almost want to think about mountain peaks. If there are multiple, distinct mountain peaks, then we say the data is multi-modal. If there is only one distinct peak, then the data is uni-modal. Not all data has a distinct peak.
- Uniform data occurs if the frequency of each interval is about the same. This will result in a flat looking histogram.
- A very important shape in statistics is the bell-curve (the shape in the first row, second column). This shape is symmetric, uni-modal and looks like a bell. If data has this shape (and satisfies a few other properties that will be discussed in Chapter 5), we call this data normal.

Here are some examples of different shapes of data:
Various shapes that data can have

Here are some examples of possible shapes that data can take



The above is provided to give you some ideas on how to describe the shape of data. But not all data sets have a nice shape that fits into one of the above. Sometimes the data can only be described as non-symmetric.

How NOT to Lie with Statistics

It is important to remember that the very reason we develop a variety of methods to present data is to develop insights into the subject of what the observations represent. We want to get a "sense" of the data. Are the observations all very much alike or are they spread across a wide range of values, are they bunched at one end of the spectrum or are they distributed evenly and so on. We are trying to get a visual picture of the numerical data. Shortly we will develop formal mathematical measures of the data, but our visual graphical presentation can say much. It can, unfortunately, also say much that is distracting, confusing and simply wrong in terms of the impression the visual leaves. Many years ago Darrell Huff wrote the book *How to Lie with Statistics*. It has been through 25 plus printings and sold more than one and one-half million copies. His perspective was a harsh one and used many actual examples that were designed to mislead. He wanted to make people aware of such deception, but perhaps more importantly to educate so that others do not make the same errors inadvertently.

Again, the goal is to enlighten with visuals that tell the story of the data. Pie charts have a number of common problems when used to convey the message of the data. Too many pieces of the pie overwhelm the reader. More than perhaps five or six categories ought to give an idea of the relative importance of each piece. This is after all the goal of a pie chart, what subset matters most relative to the others. If there are more components than this then perhaps an alternative approach would be better or perhaps some can be consolidated into an "other" category. Pie charts cannot show changes over time, although we see this attempted all too often. In federal, state, and city finance documents pie charts are often presented to show the components of revenue available to the governing body for appropriation: income tax, sales tax motor vehicle taxes and so on. In and of itself this is interesting information and can be nicely done with a pie chart. The error occurs when two years are set side-by-side. Because the total revenues change year to year, but the size of the pie is fixed, no real information is provided and the relative size of each piece of the pie cannot be meaningfully compared.

Histograms can be very helpful in understanding the data. Properly presented, they can be a quick visual way to present probabilities of different categories by the simple visual of comparing relative areas in each category. Here the error, purposeful or not, is to vary the width of the categories. This of course makes comparison to the other categories impossible. It does embellish the importance of the category with the expanded width because it has a greater area, inappropriately, and thus visually "says" that that category has a higher probability of occurrence.

Changing the units of measurement of the axis can smooth out a drop or accentuate one. If you want to show large changes, then measure the variable in small units, penny rather than thousands of dollars. And of course to continue the fraud, be sure that the axis does not begin at zero, zero. If it begins at zero, zero, then it becomes apparent that the axis has been manipulated.

Again, the goal of descriptive statistics is to convey meaningful visuals that tell the story of the data. Purposeful manipulation is fraud and unethical at the worst, but even at its best, making these type of errors will lead to confusion on the part of the analysis.

References

Burbary, Ken. *Facebook Demographics Revisited – 2001 Statistics,* 2011. Available online at http://www.kenburbary.com/2011/03/facebook-demographics-revisited-2011-statistics-2/ (accessed August 21, 2013).

"9th Annual AP Report to the Nation." CollegeBoard, 2013. Available online at http://apreport.collegeboard.org/goals-and-findings/promoting-equity (accessed September 13, 2013).

"Overweight and Obesity: Adult Obesity Facts." Centers for Disease Control and Prevention. Available online at http://www.cdc.gov/obesity/data/adult.html (accessed September 13, 2013).

Data on annual homicides in Detroit, 1961–73, from Gunst & Mason's book 'Regression Analysis and its Application', Marcel Dekker

"Timeline: Guide to the U.S. Presidents: Information on every president's birthplace, political party, term of office, and more." Scholastic, 2013. Available online at http://www.scholastic.com/teachers/article/timeline-guide-us-presidents (accessed April 3, 2013).

"Presidents." Fact Monster. Pearson Education, 2007. Available online at http://www.factmonster.com/ipka/A0194030.html (accessed April 3, 2013).

"Food Security Statistics." Food and Agriculture Organization of the United Nations. Available online at http://www.fao.org/economic/ess/ess-fs/en/ (accessed April 3, 2013).

"Consumer Price Index." United States Department of Labor: Bureau of Labor Statistics. Available online at http://data.bls.gov/pdq/SurveyOutputServlet (accessed April 3, 2013).

"CO2 emissions (kt)." The World Bank, 2013. Available online at http://databank.worldbank.org/data/home.aspx (accessed April 3, 2013).

"Births Time Series Data." General Register Office For Scotland, 2013. Available online at http://www.gro-scotland.gov.uk/statistics/theme/vital-events/births/time-series.html (accessed April 3, 2013).

"Demographics: Children under the age of 5 years underweight." Indexmundi. Available online at http://www.indexmundi.com/g/r.aspx?t=50&v=2224&aml=en (accessed April 3, 2013).

Gunst, Richard, Robert Mason. *Regression Analysis and Its Application: A Data-Oriented Approach*. CRC Press: 1980.

"Overweight and Obesity: Adult Obesity Facts." Centers for Disease Control and Prevention. Available online at http://www.cdc.gov/obesity/data/adult.html (accessed September 13, 2013).

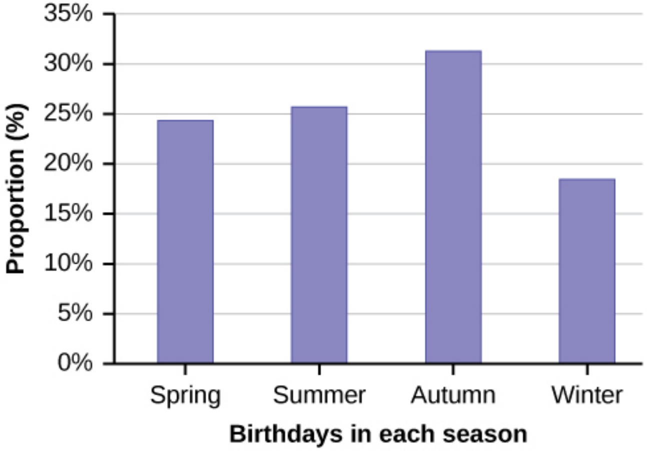
Chapter Review

A **bar graph** is a chart that uses either horizontal or vertical bars to show comparisons among categories. One axis of the chart shows the specific categories being compared, and the other axis represents a discrete value. Some bar graphs present bars clustered in groups of more than one (grouped bar graphs), and others show the bars divided into subparts to show cumulative effect (stacked bar graphs). Bar graphs are especially useful when categorical data is being used, but they can also be used for quantitative discrete data.

A **histogram** is a graphic version of a frequency distribution. The graph consists of bars of equal width drawn adjacent to each other. The horizontal scale represents classes of quantitative data values and the vertical scale represents frequencies. The heights of the bars correspond to frequency values. Histograms are typically used for large, continuous, quantitative data sets.

The students in Ms. Ramirez’s math class have birthdays in each of the four seasons. [link] shows the four seasons, the number of students who have birthdays in each season, and the percentage (%) of students in each group. Construct a bar graph showing the percentage of students in each group.

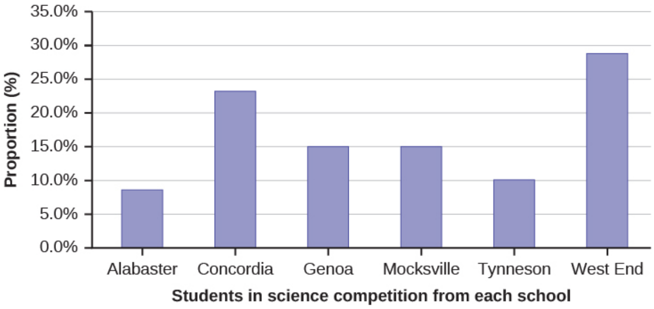
Seasons	Number of students	Proportion of population
Spring	8	24%
Summer	9	26%
Autumn	11	32%
Winter	6	18%



David County has six high schools. Each school sent students to participate in a county-wide science competition. [link] shows the percentage breakdown of competitors from each school, and the percentage of the entire student population of the county that goes to each school. Construct a bar graph that shows the county-wide population percentage of students at each school.

High School	Science competition population	Overall student population
Alabaster	28.9%	8.6%
Concordia	7.6%	23.2%

Genoa	12.1%	15.0%
Mocksville	18.5%	14.3%
Tynneson	24.2%	10.1%
West End	8.7%	28.8%



Construct a histogram for the following:

1. **Pulse Rates for Women**	**Frequency**
60–69	12
70–79	14
80–89	11
90–99	1
100–109	1
110–119	0

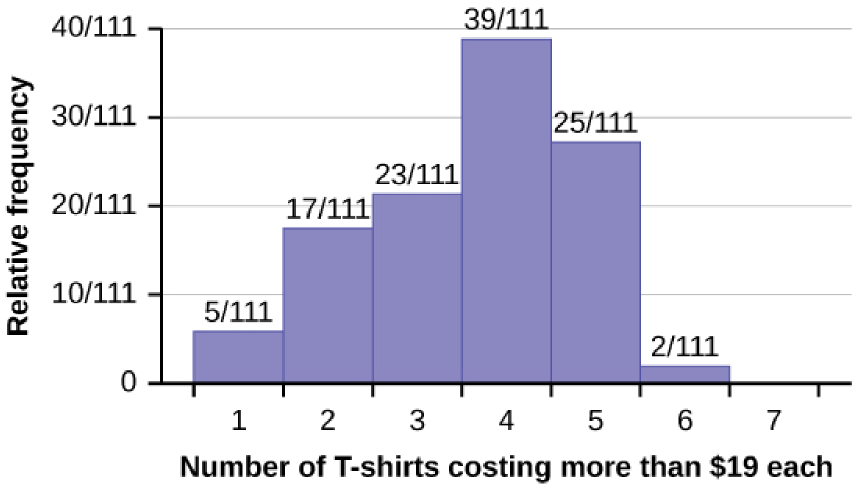
120–129	1	

2. **Actual Speed in a 30 MPH Zone**	**Frequency**	
42–45	25	
46–49	14	
50–53	7	
54–57	3	
58–61	1	

3. **Tar (mg) in Nonfiltered Cigarettes**	**Frequency**	
10–13	1	
14–17	0	
18–21	15	
22–25	7	
26–29	2	

Homework

Use the following information to answer the next two exercises: Suppose one hundred eleven people who shopped in a special t-shirt store were asked the number of t-shirts they own costing more than $19 each.



The percentage of people who own at most three t-shirts costing more than $19 each is approximately:

1. 21
2. 59
3. 41
4. Cannot be determined

c

If the data were collected by asking the first 111 people who entered the store, then the type of sampling is:

1. cluster
2. simple random
3. stratified
4. convenience

d

Glossary

Frequency
: the number of times a value of the data occurs

Histogram
: a graphical representation in *x*-*y* form of the distribution of data in a data set; *x* represents the data and *y* represents the frequency, or relative frequency. The graph consists of contiguous rectangles.

Relative Frequency
: the ratio of the number of times a value of the data occurs in the set of all outcomes to the

number of all outcomes

Introduction

Measures of location help us to understand where data values are located relative to other data values. We've already seen a measure of location - the median. It tells us what data value is in the middle of the data set. The most common measure of position is a **percentile** . Percentiles divide ordered data into hundredths. To score in the 90th percentile of an exam does not mean, necessarily, that you received 90% on a test. It means that 90% of test scores are the same or less than your score and 10% of the test scores are the same or greater than your test score. The median is the 50th percentile

A special type of percentile are called **quartiles**. Quartiles divide ordered data into quarters. The first quartile, Q_1, is the same as the 25th percentile, and the third quartile, Q_3, is the same as the 75th percentile. The median, M, is called both the second quartile and the 50th percentile.

A visual representation of measures of location is called a **box plot**.

In this section, we will learn how to find quartiles and use those quartiles to find the interquartile range and outliers. Then we will visually represent this information on a box plot. Unlike histograms and bar graphs, box plots require the use of numerical summaries. Thus the box plot is a representation that combines both visual and numerical summaries of the data.

Measures of location

As described in the introduction, a common measure of location are percentiles. Percentiles are useful for comparing values. For this reason, universities and colleges use percentiles extensively. One instance in which colleges and universities use percentiles is when SAT results are used to determine a minimum testing score that will be used as an acceptance factor. For example, suppose Duke accepts SAT scores at or above the 75th percentile. That translates into a score of at least 1220.

Percentiles are mostly used with very large populations. Therefore, if you were to say that 90% of the test scores are less (and not the same or less) than your score, it would be acceptable because removing one particular data value is not significant.

The **median** is a number that measures the "center" of the data. You can think of the median as the "middle value," but it does not actually have to be

one of the observed values. It is a number that separates ordered data into halves. Half the values are the same number or smaller than the median, and half the values are the same number or larger. For example, consider the following data.
1; 11.5; 6; 7.2; 4; 8; 9; 10; 6.8; 8.3; 2; 2; 10; 1
Ordered from smallest to largest:
1; 1; 2; 2; 4; 6; 6.8; 7.2; 8; 8.3; 9; 10; 10; 11.5

Since there are 14 observations, the median is between the seventh value, 6.8, and the eighth value, 7.2. To find the median, add the two values together and divide by two.
$6.8 + 7.22 = 7$

The median is seven. Half of the values are smaller than seven and half of the values are larger than seven.

Quartiles are numbers that separate the data into quarters. Quartiles may or may not be part of the data. To find the quartiles, first find the median or second quartile. The first quartile, *Q*1, is the middle value of the lower half of the data, and the third quartile, *Q*3, is the middle value, or median, of the upper half of the data. To get the idea, consider the same data set:
1; 1; 2; 2; 4; 6; 6.8; 7.2; 8; 8.3; 9; 10; 10; 11.5

Quartiles are numbers that separate the data into quarters. Quartiles may or may not be part of the

data. To find the quartiles, first find the median or second quartile. The first quartile, *Q*1, is the middle value of the lower half of the data, and the third quartile, *Q*3, is the middle value, or median, of the upper half of the data. To get the idea, consider the same data set:
1; 1; 2; 2; 4; 6; 6.8; 7.2; 8; 8.3; 9; 10; 10; 11.5

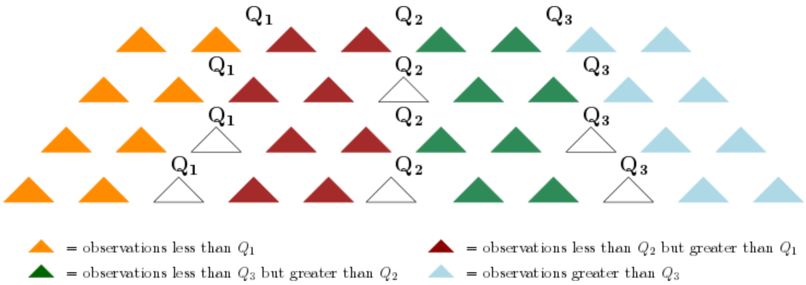
The median or **second quartile** is seven. The lower half of the data are 1, 1, 2, 2, 4, 6, 6.8. The middle value of the lower half is two.
1; 1; 2; 2; 4; 6; 6.8

The number two, which is part of the data, is the **first quartile**. One-fourth of the entire sets of values are the same as or less than two and three-fourths of the values are more than two.

The upper half of the data is 7.2, 8, 8.3, 9, 10, 10, 11.5. The middle value of the upper half is nine.

The **third quartile**, *Q*3, is nine. Three-fourths (75%) of the ordered data set are less than nine. One-fourth (25%) of the ordered data set are greater than nine. The third quartile is part of the data set in this example.

Possible Quartile Positions


Q1 Q2 Q3
Q1 Q2 Q3
Q1 Q2 Q3
Q1 Q2 Q3
= observations less than Q_1
= observations less than Q_2 but greater than Q_1
= observations less than Q_3 but greater than Q_2
= observations greater than Q_3

As mentioned in the previous section, the **interquartile range** is a measure of variation. It is a number that indicates the spread of the middle half or the middle 50% of the data. It is the difference between the third quartile (Q_3) and the first quartile (Q_1).

$IQR = Q_3 – Q_1$

The *IQR* can help to determine potential **outliers**. **A value is suspected to be a potential outlier if it is less than (1.5)(*IQR*) below the first quartile or more than (1.5)(*IQR*) above the third quartile.** Potential outliers always require further investigation.

NOTE

A potential outlier is a data point that is significantly different from the other data points. These special data points may be errors or some kind of abnormality or they may be a key to understanding the data.

For the following 13 real estate prices, calculate the *IQR* and determine if any prices are potential outliers. Prices are in dollars. 389,950; 230,500; 158,000; 479,000; 639,000; 114,950; 5,500,000; 387,000; 659,000; 529,000; 575,000; 488,800; 1,095,000

Order the data from smallest to largest. 114,950; 158,000; 230,500; 387,000; 389,950; 479,000; 488,800; 529,000; 575,000; 639,000; 659,000; 1,095,000; 5,500,000

$M = 488{,}800$

$Q_1 = 230{,}500 + 387{,}000\ 2 = 308{,}750$

$Q_3 = 639{,}000 + 659{,}000\ 2 = 649{,}000$

$IQR = 649{,}000 – 308{,}750 = 340{,}250$

$(1.5)(IQR) = (1.5)(340{,}250) = 510{,}375$

1.5(IQR) less than the first quartile: $Q_1 – (1.5)(IQR) = 308{,}750 – 510{,}375 = –201{,}625$

1.5(IQR) more than the first quartile:$Q_3 + (1.5)(IQR) = 649{,}000 + 510{,}375 = 1{,}159{,}375$

No house price is less than –201,625. However, 5,500,000 is more than 1,159,375. Therefore, 5,500,000 is a potential **outlier**.

For the two data sets in the test scores example, find the following:

1. The interquartile range. Compare the two interquartile ranges.
2. Any outliers in either set.

The five number summary for the day and night classes is

	Minimum	Q_1	Median	Q_3	Maximum
Day	32	56	74.5	82.5	99
Night	25.5	78	81	89	98

1. The IQR for the day group is $Q_3 - Q_1 = 82.5 - 56 = 26.5$

   The IQR for the night group is $Q_3 - Q_1 =$

$89 – 78 = 11$

The interquartile range (the spread or variability) for the day class is larger than the night class *IQR*. This suggests more variation will be found in the day class's class test scores.

2. Day class outliers are found using the IQR times 1.5 rule. So,

   - $Q_1 - IQR(1.5) = 56 – 26.5(1.5) = 16.25$
   - $Q_3 + IQR(1.5) = 82.5 + 26.5(1.5) = 122.25$

   Since the minimum and maximum values for the day class are greater than 16.25 and less than 122.25, there are no outliers.

   Night class outliers are calculated as:

   - $Q_1 – IQR\ (1.5) = 78 – 11(1.5) = 61.5$
   - $Q_3 + \mathrm{IQR}(1.5) = 89 + 11(1.5) = 105.5$

   For this class, any test score less than 61.5 is an outlier. Therefore, the scores of 45 and 25.5 are outliers. Since no test score is greater than 105.5, there is no upper end outlier.

Interpreting Percentiles, Quartiles, and Median

A percentile indicates the relative standing of a data value when data are sorted into numerical order from smallest to largest. Percentages of data values are less than or equal to the pth percentile. For example, 15% of data values are less than or equal to the 15th percentile.

- Low percentiles always correspond to lower data values.
- High percentiles always correspond to higher data values.

A percentile may or may not correspond to a value judgment about whether it is "good" or "bad." The interpretation of whether a certain percentile is "good" or "bad" depends on the context of the situation to which the data applies. In some situations, a low percentile would be considered "good;" in other contexts a high percentile might be considered "good". In many situations, there is no value judgment that applies.

Understanding how to interpret percentiles properly is important not only when describing data, but also when calculating probabilities in later chapters of this text.

Guideline

When writing the interpretation of a percentile in the context of the given data, the sentence should contain the following information.

- information about the context of the situation being considered
- the data value (value of the variable) that represents the percentile
- the percent of individuals or items with data values below the percentile
- the percent of individuals or items with data values above the percentile.

On a timed math test, the first quartile for time it took to finish the exam was 35 minutes. Interpret the first quartile in the context of this situation.

- Twenty-five percent of students finished the exam in 35 minutes or less.
- Seventy-five percent of students finished the exam in 35 minutes or more.
- A low percentile could be considered good, as finishing more quickly on a timed exam is desirable. (If you take too long, you might not be able to finish.)

On a 20 question math test, the 70th percentile for number of correct answers was 16. Interpret the 70th percentile in the context of this situation.

- Seventy percent of students answered 16 or fewer questions correctly.
- Thirty percent of students answered 16 or more questions correctly.
- A higher percentile could be considered good, as answering more questions correctly is desirable.

Try It

On a 60 point written assignment, the 80th percentile for the number of points earned was 49. Interpret the 80th percentile in the context of this situation.

Eighty percent of students earned 49 points or fewer. Twenty percent of students earned 49 or more points. A higher percentile is good because getting more points on an assignment is desirable.

At a community college, it was found that the 30th percentile of credit units that students are enrolled for is seven units. Interpret the 30th percentile in the context of this situation.

- Thirty percent of students are enrolled in seven or fewer credit units.
- Seventy percent of students are enrolled in seven or more credit units.
- In this example, there is no "good" or "bad" value judgment associated with a higher or lower percentile. Students attend community college for varied reasons and needs, and their course load varies according to their needs.

Outliers

Above the idea of potential outliers were discussed. This section will look more in depth at how to find outliers and how to categorize them.

Quartiles can also be used to determine if there are any outliers in a data set. To determine if there are outliers, we need to first calculate the **inner and outer fences**. The fences define the boundary between a “normal” data value and an “abnormal” data value (or outlier). Any data values that fall between the inner fences are normal data values. **Any data values that fall outside the inner fences are considered outliers**.

The fences are calculated as follows:

The inner fences are Q_1 - $IQR(1.5)$ and $Q_3 + IQR(1.5)$.

The outer fences are Q_1 - $IQR(3)$ and $Q_3 + IQR(3)$.

A mild outlier is any data value between the inner and outer fences.

An extreme outlier is any data value to the extreme of the outer fence.

Finding outliers

Sharpe Middle School is applying for a grant that will be used to add fitness equipment to the gym.

The principal surveyed 15 anonymous students to determine how many minutes a day the students spend exercising. The results from the 15 anonymous students are shown.

0 minutes; 40 minutes; 60 minutes; 30 minutes; 60 minutes 10 minutes; 45 minutes; 30 minutes; 300 minutes; 90 minutes; 30 minutes; 120 minutes; 60 minutes; 0 minutes; 20 minutes

The five-number summary is determined to be: Min = 0; Q1 = 20; Med = 40; Q3 = 60; Max = 300.

Are there any students who are exercising significantly more or less than the other students?

To answer this question, we have to determine if there are any outliers.

To do this, determine the inner fences.

The IQR is 60-20 = 40.

The lower inner fence is $Q_1 - IQR(1.5) = 20 – 40(1.5) = -40$\$ and the upper inner fence is $Q_3 + IQR(1.5) = 60 + 40(1.5) = 120$\$. Thus, any student who exercises between -40 minutes and 120 minutes is exercising a “normal” amount of time (relative to the rest of the students). Since someone can’t exercise -40 minutes, this is really 0 minutes to 120 minutes. Therefore, 300 minutes appears to be an outlier. But is it a mild outlier or an extreme outlier?

To determine if it is mild or extreme, we need to calculate the outer fence. We only need the upper outer fence as there are no low outliers (no one exercised for less than -40 minutes). The upper outer fence is $Q + IQR(3) = 60 + 40(3) = 180$\$.

If the potential outlier is between 120 and 180 minutes, then it is a mild outlier (as it is between the upper inner and outer fences). If it is more than 180 minutes, then it is an extreme outlier. In this case, 300 minutes is an extreme outlier. This means that this student is exercising way more than the rest of their classmates!

Box Plots

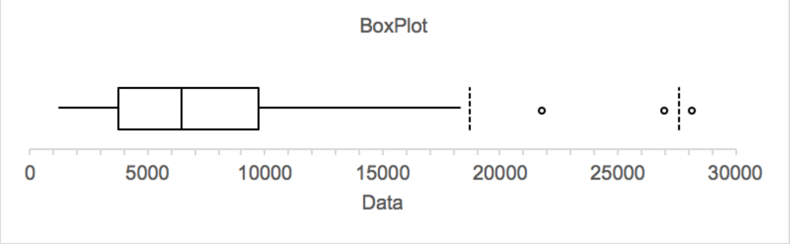
Box plots (also called **box-and-whisker plots** or **box-whisker plots**) give a good graphical image of the concentration of the data. They also show how far the extreme values are from most of the data.

To construct a box plot, use a horizontal or vertical number line and a rectangular box. The smallest and largest data values label the endpoints of the axis. The first quartile marks one end of the box and the third quartile marks the other end of the box. Approximately **the middle 50 percent of the data fall inside the box.** The "whiskers" extend from the ends of the box to the smallest and largest data values. The median or second quartile can be between the first and third quartiles, or it can be one, or the other, or both. The box plot gives a good, quick picture of the data.

A box plot is constructed from the five-number summary (the minimum value, the first quartile, the median, the third quartile, and the maximum value) and, if there are outliers, the fences. We use these values to compare how close other data values are to them.

Example of a box plot

This is an example of a box plot. The box is in the middle and represents 50% of the data. The circles on the right represent outliers and the dashed lines the fences. The outliers at approximately 22000 and 27000 are mild outliers, while the outlier at approximately 28500 is an extreme outlier.



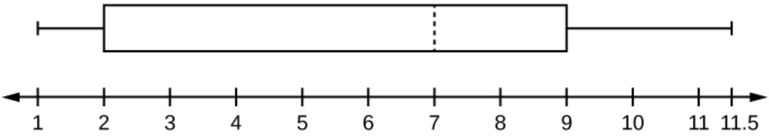
To construct a box plot, use a horizontal or vertical number line and a rectangular box. The smallest and largest data values label the endpoints of the axis. The first quartile marks one end of the box and the third quartile marks the other end of the box. The median is represented by a line inside the box. The middle 50 percent of the data fall inside the box and the length of the box is the interquartile range.

The "whiskers" extend from the ends of the box to the first data values inside the fences. If there are no outliers, this would be minimum and maximum values. The outliers are represented by asterisks or dots and fall either between the inner and outer fences (**mild outlier**) or outside the outer fences (**extreme outlier**).

Consider, again, this dataset.

1 1 2 2 4 6 6.8 7.2 8 8.3 9 10 10 11.5

From the work done above, we know the five number summary is 1, 2, 7, 9, 11.5. The IQR is 9-2 = 7. IQR(1.5) is 7*1.5 = 10.5. The lower inner fence is Q1-IQR(1.5) = 2-10.5=-8.5 and the upper inner fence is Q3+IQR(1.5)=9+10.5 = 19.5. Since no data values are smaller than -8.5 or larger than 19.5, there are no outliers in the data set.



The two whiskers extend from the first quartile to the smallest value and from the third quartile to the largest value. The median is shown with a dashed line.

NOTE

It is important to start a box plot with a **scaled number line**. Otherwise the box plot may not be useful.

The following data are the heights of 40 students (in inches) in a statistics class.
59; 60; 61; 62; 62; 63; 63; 64; 64; 64; 65; 65; 65; 65; 65; 65; 65; 65; 65; 66; 66; 67; 67; 68; 68; 69; 70; 70; 70; 70; 70; 71; 71; 72; 72; 73; 74; 74; 75; 77
Take this data and input it into Excel. Use the "Text to Columns" function in the "Data" menu to separate the data into separate columns. Then copy the data, but when you paste it, use paste special to "Transpose" the data so it is all in one column.
Now use whatever software you are using to find the five-number summary.

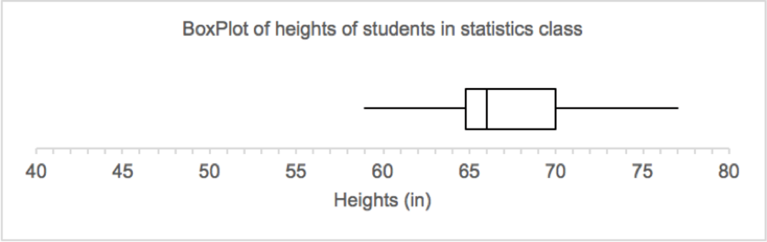
- Minimum value = 59
- Q1: First quartile = 64.75
- Q2: Second quartile or median = 66
- Q3: Third quartile = 70
- Maximum value = 77

Are there outliers? The IQR is 70-64.75 = 5.25.
IQR(1.5) = 7.875 (don't round until the end)
The lower inner fence is Q1 - IQR(1.5) = 64.75-7.875 = 56.875. Since the minimum value is 59, there are no lower outliers.
The upper inner fence is Q3 + IQR(1.5) = 70+7.875 = 77.875. Since the maximum value is

77, there are no upper outliers.

You can also use your computer program to create a box plot for the data.

Box plot of height of 40 students



The titles and labels for a box plot follow the same rules as they do for a histogram or a bar graph.

What does the box plot tell us?

- Each quarter has approximately 25% of the data.
- The spreads of the four quarters are 64.75 – 59 = 5.75 (first quarter), 66 – 64.75 = 1.25 (second quarter), 70 – 66= 4 (third quarter), and 77 – 70 = 7 (fourth quarter). So, the second quarter has the smallest spread and the fourth quarter has the largest spread.
- Range = maximum value – the minimum value = 77 – 59 = 18, which means that from the shortest to the tallest student there is

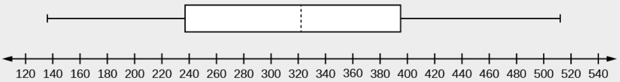
a difference of 18 inches.

- Interquartile Range: IQR = third quartile - first quartile = 70 – 64.75 = 5.25, which means that the middle 50% (middle half) of the data has a range of 5.25 inches. This also means the length of the box is 5.25.

Try It

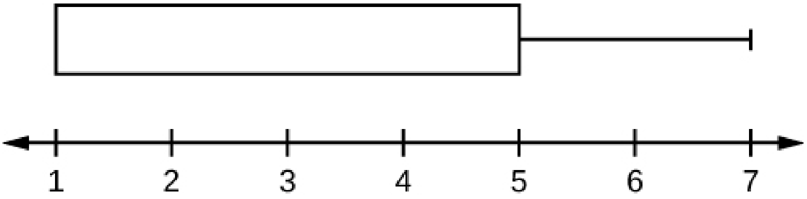
The following data are the number of pages in 40 books on a shelf. Construct a box plot using computer software, and state the interquartile range.

136 140 178 190 205 215 217 218 232 234
240 255 270 275 290 301 303 315 317 318
326 333 343 349 360 369 377 388 391 392
398 400 402 405 408 422 429 450 475 512



$IQR = 158$

For some sets of data, some of the largest value, smallest value, first quartile, median, and third quartile may be the same. For instance, you might have a data set in which the median and the third quartile are the same. In this case, the diagram would not have a dotted line inside the box displaying the median. The right side of the box would display both the third quartile and the median. For example, if the smallest value and the first quartile were both one, the median and the third quartile were both five, and the largest value was seven, the box plot would look like:



In this case, at least 25% of the values are equal to one. Twenty-five percent of the values are between one and five, inclusive. At least 25% of the values are equal to five. The top 25% of the values fall between five and seven, inclusive.

Test scores for a college statistics class held during the day are:
99 56 78 55.5 32 90 80 81 56 59 45 77 84.5 84 70 72 68 32 79 90
Test scores for a college statistics class held during the evening are:
98 78 68 83 81 89 88 76 65 45 98 90 80 84.5 85

79 78 98 90 79 81 25.5

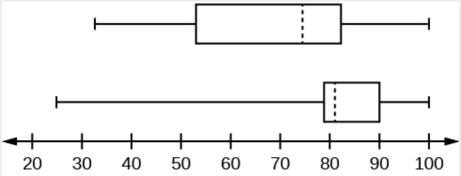
1. Find the smallest and largest values, the median, and the first and third quartile for the day class.
2. Find the smallest and largest values, the median, and the first and third quartile for the night class.
3. For each data set, what percentage of the data is between the smallest value and the first quartile? the first quartile and the median? the median and the third quartile? the third quartile and the largest value? What percentage of the data is between the first quartile and the largest value?
4. Create a box plot for each set of data. Use one number line for both box plots.
5. Which box plot has the widest spread for the middle 50% of the data (the data between the first and third quartiles)? What does this mean for that set of data in comparison to the other set of data?

1.
    - Min = 32
    - $Q_1 = 56$
    - $M = 74.5$
    - $Q_3 = 82.5$
    - Max = 99

7.
- Min = 25.5
- Q_1 = 78
- M = 81
- Q_3 = 89
- Max = 98

13. Day class: There are six data values ranging from 32 to 56: 30%. There are six data values ranging from 56 to 74.5: 30%. There are five data values ranging from 74.5 to 82.5: 25%. There are five data values ranging from 82.5 to 99: 25%. There are 16 data values between the first quartile, 56, and the largest value, 99: 75%. Night class:

14.



15. The first data set has the wider spread for the middle 50% of the data. The *IQR* for the first data set is greater than the *IQR* for the second set. This means that there is more variability in the middle 50% of the first data set.

Try It

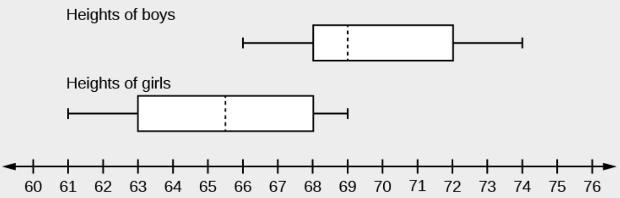
The following data set shows the heights in inches for the boys in a class of 40 students.

66; 66; 67; 67; 68; 68; 68; 68; 68; 69; 69; 69; 70; 71; 72; 72; 72; 73; 73; 74

The following data set shows the heights in inches for the girls in a class of 40 students.

61; 61; 62; 62; 63; 63; 63; 65; 65; 65; 66; 66; 66; 67; 68; 68; 68; 69; 69; 69

Construct a box plot using computer software for each data set, and state which box plot has the wider spread for the middle 50% of the data.



IQR for the boys = 4

IQR for the girls = 5

The box plot for the heights of the girls has the wider spread for the middle 50% of the data.

References

Data from *West Magazine*.

Cauchon, Dennis, Paul Overberg. “Census data shows minorities now a majority of U.S. births.” USA Today, 2012. Available online at http://usatoday30.usatoday.com/news/nation/story/2012-05-17/minority-birthscensus/55029100/1 (accessed April 3, 2013).

Data from the United States Department of Commerce: United States Census Bureau. Available online at http://www.census.gov/ (accessed April 3, 2013).

“1990 Census.” United States Department of Commerce: United States Census Bureau. Available online at http://www.census.gov/main/www/cen1990.html (accessed April 3, 2013).

Data from *San Jose Mercury News*.

Data from *Time Magazine*; survey by Yankelovich Partners, Inc.

Chapter Review

The values that divide a rank-ordered set of data into 100 equal parts are called percentiles.

Percentiles are used to compare and interpret data. For example, an observation at the 50th percentile would be greater than 50 percent of the other obeservations in the set. Quartiles divide data into quarters. The first quartile (Q_1) is the 25th percentile,the second quartile (Q_2 or median) is 50th percentile, and the third quartile (Q_3) is the the 75th percentile. The interquartile range, or *IQR*, is the range of the middle 50 percent of the data values. The *IQR* is found by subtracting Q_1 from Q_3, and can help determine outliers by using the following two expressions.

- $Q_3 + IQR(1.5)$
- $Q_1 – IQR(1.5)$

Box plots are a type of graph that can help visually organize data. To graph a box plot the following data points must be calculated: the minimum value, the first quartile, the median, the third quartile, and the maximum value. Once the box plot is graphed, you can display and compare distributions of data.

On an exam, would it be more desirable to earn a grade with a high or low percentile? Explain.

It is better to earn a grade in a high percentile as that means that you have done better on the exam relative to your classmates.

Mina is waiting in line at the Department of Motor Vehicles (DMV). Her wait time of 32 minutes is the 85th percentile of wait times. Is that good or bad? Write a sentence interpreting the 85th percentile in the context of this situation.

When waiting in line at the DMV, the 85th percentile would be a long wait time compared to the other people waiting. 85% of people had shorter wait times than Mina. In this context, Mina would prefer a wait time corresponding to a lower percentile. 85% of people at the DMV waited 32 minutes or less. 15% of people at the DMV waited 32 minutes or longer.

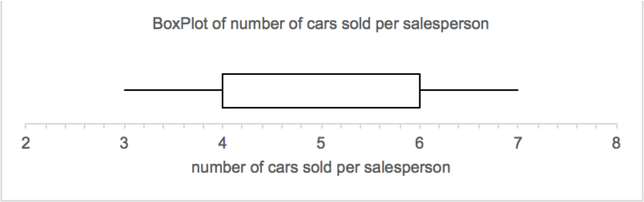
In a study collecting data about the repair costs of damage to automobiles in a certain type of crash tests, a certain model of car had $1,700 in damage and was in the 90th percentile. Should the manufacturer and the consumer be pleased or upset by this result? Explain and write a sentence that interprets the 90th percentile in the context of this problem.

The manufacturer and the consumer would be upset. This is a large repair cost for the damages, compared to the other cars in the sample. INTERPRETATION: 90% of the crash tested cars had damage repair costs of $1700 or less; only 10% had damage repair costs of $1700 or more.

Suppose that you are buying a house. You and your realtor have determined that the most expensive house you can afford is the 34th percentile. The 34th percentile of housing prices is $240,000 in the town you want to move to. In this town, can you afford 34% of the houses or 66% of the houses?

You can afford 34% of houses. 66% of the houses are too expensive for your budget. INTERPRETATION: 34% of houses cost $240,000 or less. 66% of houses cost $240,000 or more.

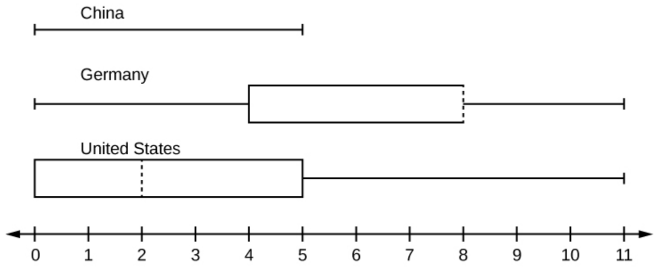
Sixty-five randomly selected car salespersons were asked the number of cars they generally sell in one week. Fourteen people answered that they generally sell three cars; nineteen generally sell four cars; twelve generally sell five cars; nine generally sell six cars; eleven generally sell seven cars. Construct a box plot for this data.



Looking at your box plot in the exercise above, does it appear that the data are concentrated together, spread out evenly, or concentrated in some areas, but not in others? How can you tell?

More than 25% of salespersons sell four cars in a typical week. You can see this concentration in the box plot because the first quartile is equal to the median. The top 25% and the bottom 25% are spread out evenly; the whiskers have the same length.

In a survey of 20-year-olds in China, Germany, and the United States, people were asked the number of foreign countries they had visited in their lifetime. The following box plots display the results.



1. In complete sentences, describe what the shape of each box plot implies about the distribution of the data collected.
2. Have more Americans or more Germans surveyed been to over eight foreign countries?
3. Compare the three box plots. What do they imply about the foreign travel of 20-year-old residents of the three countries when compared to each other?

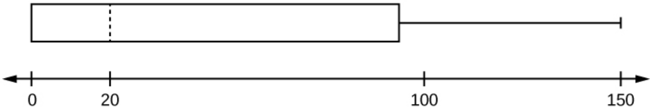
---

1. The shape of China suggests that either every person they surveyed except one either visited 0 foreign countries or 5 foreign countries. For example, if 30 people were interviewed in China, 29 of them have visited no foreign country and one of them has visited 5 foreign countries OR 29 of them have visited 5 foreign countries and one of them has visited no foreign countries. It is unclear which way it is going in the box plot. In Germany, 50% of those surveyed have visited 8 or less countries. Based on the position of the

median, this suggests that there are many people in the survey who have visited eight countries. This suggests the distribution will have a peak at 8 and will be non-symmetric. In the USA, 50% of those surveyed have visited 2 or less countries. As there are no whiskers, this suggests that 25% of the Americans surveyed have visited no foreign countries which suggest a skew to the right for the distribution.

2. 25% of Germans surveyed have been to more than 8 foreign countries. It is unclear what the percentage is for Americans but it is less than 25%. Therefore, Germany.
3. Germans in the survey have visited far more countries that Americans and the Chinese in the survey. China has the least foreign travel.

Given the following box plot, answer the questions.



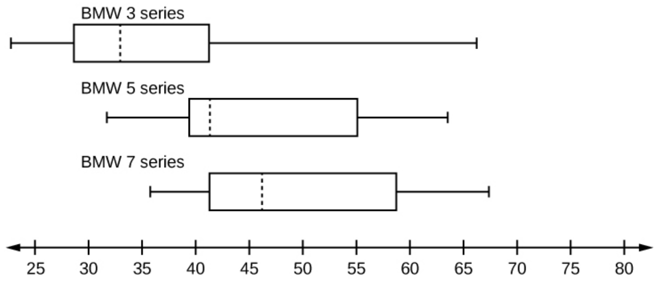
1. Think of an example (in words) where the data might fit into the above box plot. In 2–5 sentences, write down the example.

2. What does it mean to have the first and second quartiles so close together, while the second to third quartiles are far apart?

---

1. Answers will vary. Possible answer: State University conducted a survey to see how involved its students are in community service. The box plot shows the number of community service hours logged by participants over the past year.
2. Because the first and second quartiles are close, the data in this quarter is very similar. There is not much variation in the values. The data in the third quarter is much more variable, or spread out. This is clear because the second quartile is so far away from the third quartile.

A survey was conducted of 130 purchasers of new BMW 3 series cars, 130 purchasers of new BMW 5 series cars, and 130 purchasers of new BMW 7 series cars. In it, people were asked the age they were when they purchased their car. The following box plots display the results.



1. In complete sentences, describe what the shape of each box plot implies about the distribution of the data collected for that car series.
2. Which group is most likely to have an outlier? Explain how you determined that.
3. Compare the three box plots. What do they imply about the age of purchasing a BMW from the series when compared to each other?
4. Look at the BMW 5 series. Which quarter has the smallest spread of data? What is the spread?
5. Look at the BMW 5 series. Which quarter has the largest spread of data? What is the spread?
6. Look at the BMW 5 series. Estimate the interquartile range (IQR).
7. Look at the BMW 5 series. Are there more data in the interval 31 to 38 or in the interval 45 to 55? How do you know this?
8. Look at the BMW 5 series. Which interval has the fewest data in it? How do you know this?

1. 31–35
2. 38–41
3. 41–64

---

1. Each box plot is spread out more in the greater values. Each plot is skewed to the right, so the ages of the top 50% of buyers are more variable than the ages of the lower 50%.
2. The BMW 3 series is most likely to have an outlier. It has the longest whisker.
3. Comparing the median ages, younger people tend to buy the BMW 3 series, while older people tend to buy the BMW 7 series. However, this is not a rule, because there is so much variability in each data set.
4. The second quarter has the smallest spread. There seems to be only a three-year difference between the first quartile and the median.
5. The third quarter has the largest spread. There seems to be approximately a 14-year difference between the median and the third quartile.
6. $IQR \sim 17$ years
7. There is not enough information to tell. Each interval lies within a quarter, so we cannot tell exactly where the data in that quarter is concentrated.

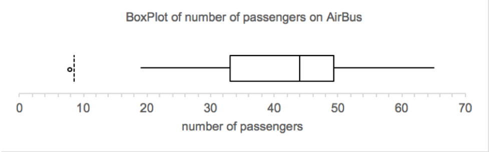
8. The interval from 31 to 35 years has the fewest data values. Twenty-five percent of the values fall in the interval 38 to 41, and 25% fall between 41 and 64. Since 25% of values fall between 31 and 38, we know that fewer than 25% fall between 31 and 35.

The following data represents the number of passengers per flight on the AirBus from Calgary to Edmonton for 24 flights.

8, 19, 22, 23, 29, 30, 34, 35, 37, 39, 41, 44, 44, 46, 46, 47, 48, 49, 50, 52, 54, 55, 61, 65

1. Generate the boxplot for this data.
2. Identify the outliers in the data. Are they low or high outliers? Are the extreme or mild outliers?
3. Interpret the outliers in the context of the question.
4. What is the IQR? Interpret it in the context of the question.
5. Which quarter of the data is the most concentrated? The least concentrated?
6. What is the five-number summary (minimum, first quartile, median, third quartile, maximum)?

1.



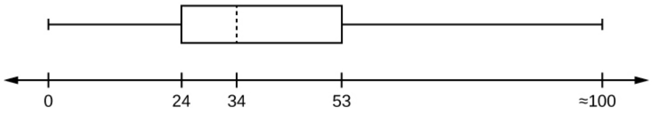
2. There is one mild low outlier of 8 passengers on a flight.
3. a) The outlier means that on this flight there were significantly fewer passengers (only 8) than there are on other similar flights.
4. The IQR is 16.25 (from 33 to 49.25). This means that 50% of the time, the number of passengers is between 33 and 49.25 on the Airbus. This gives us a sense of the amount of variation in the number of passengers.
5. The distance between the median and the third quartile (from 44 to 49.25) is the least (5.25). This means that these 25% of data values are closely packed together. While the distance between the outlier and the first quartile is the largest (25 passengers). This means that these 25% of the data values are spread out from each other.
6. a) The five-number summary is: Minimum = 8; First quartile = 33; Median = 44; Third quartile = 49.25; Maximum = 65.

Bringing It Together

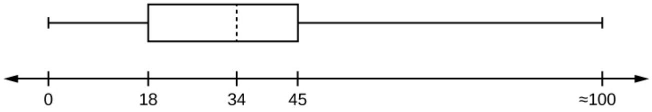
Santa Clara County, CA, has approximately 27,873 Japanese-Americans. Their ages are as follows:

Age Group	Percent of Community
0–17	18.9
18–24	8.0
25–34	22.8
35–44	15.0
45–54	13.1
55–64	11.9
65+	10.3

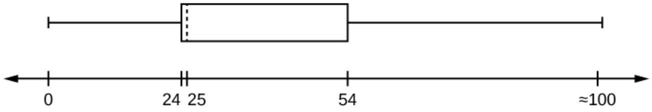
1. Construct a histogram of the Japanese-American community in Santa Clara County, CA. The bars will **not** be the same width for this example. Why not? What impact does this have on the reliability of the graph?
2. What percentage of the community is under age 35?
3. Which box plot most resembles the information above?



(a)

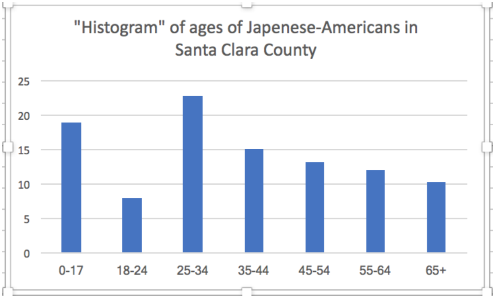


(b)



(c)

1. This is technically not a histogram as the bars aren't touching, but without the original data this is the best that I could come up with unless I drew it by hand!



2. 49.7% of the community is under the age of 35.
3. Based on the information in the table, graph (a) most closely represents the data.

Glossary

Box plot

a graph that gives a quick picture of the middle 50% of the data

First Quartile

the value that is the median of the of the lower half of the ordered data set

Frequency Polygon

looks like a line graph but uses intervals to display ranges of large amounts of data

Interval

also called a class interval; an interval represents a range of data and is used when displaying large data sets

Paired Data Set

two data sets that have a one to one relationship so that:

- both data sets are the same size, and
- each data point in one data set is matched with exactly one point from the other set.

Skewed

used to describe data that is not symmetrical; when the right side of a graph looks “chopped off” compared the left side, we say it is “skewed to the left.” When the left side of the graph looks “chopped off” compared to the right side, we say the data is “skewed to the right.” Alternatively: when the lower values of the data are more spread out, we say the data are skewed to the left. When the greater values are more spread out, the data are skewed to the right.

Introduction to Probability

Probability is a measure that is associated with how certain we are of outcomes of a particular experiment or activity. An **experiment** is a planned operation carried out under controlled conditions. If the result is not predetermined, then the experiment is said to be a **chance** experiment. Flipping one fair coin twice is an example of an experiment.

A result of an experiment is called an **outcome**. The **sample space** of an experiment is the set of all possible outcomes. Three ways to represent a sample space are: to list the possible outcomes, to create a tree diagram, or to create a Venn diagram. The uppercase letter S is used to denote the sample space. For example, if you flip one fair coin, $S = \{H, T\}$ where $H =$ heads and $T =$ tails are the outcomes.

An **event** is any combination of outcomes. Upper case letters like A and B represent events. For example, if the experiment is to flip one fair coin, event A might be getting at most one head. The probability of an event A is written $P(A)$.

The **probability** of any outcome is the **long-term relative frequency** of that outcome. **Probabilities are between zero and one, inclusive** (that is, zero and one and all numbers between these values). $P(A) = 0$ means the event A can never happen. $P(A)$

= 1 means the event A always happens. $P(A) = 0.5$ means that event A has a 50% chance of happening. For example, if you flip one fair coin repeatedly (from 20 to 2,000 to 20,000 times) the relative frequency of heads approaches 0.5 (the probability of heads).

Equally likely means that each outcome of an experiment occurs with equal probability. For example, if you toss a **fair**, six-sided die, each face (1, 2, 3, 4, 5, or 6) is as likely to occur as any other face. If you toss a fair coin, a Head (H) and a Tail (T) are equally likely to occur. If you randomly guess the answer to a true/false question on an exam, you are equally likely to select a correct answer or an incorrect answer.

To calculate the probability of an event *A* when all outcomes in the sample space are equally likely, count the number of outcomes for event A and divide by the total number of outcomes in the sample space. For example, if you toss a fair dime and a fair nickel, the sample space is $\{HH, TH, HT, TT\}$ where T = tails and H = heads. The sample space has four outcomes. A = getting one head. There are two outcomes that meet this condition $\{HT, TH\}$, so $P(A)$ = 2 4 = 0.5.

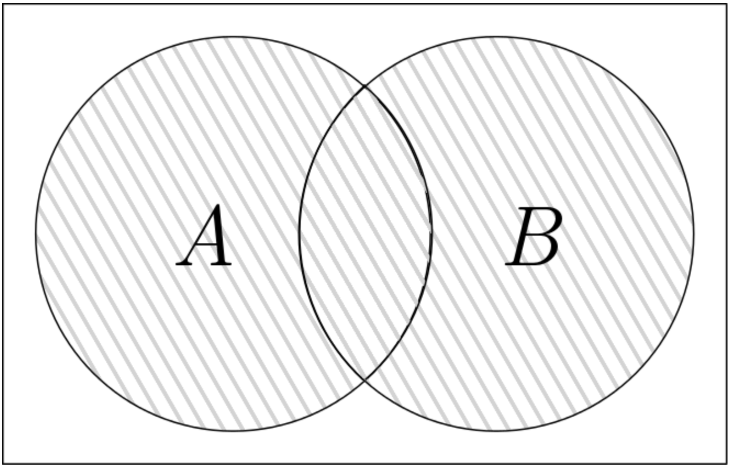
Suppose you roll one fair six-sided die, with the numbers {1, 2, 3, 4, 5, 6} on its faces. Let event E = rolling a number that is at least five. There are two outcomes {5, 6}. $P(E)$ = 2 6 . If you were to roll the die only a few times, you would not be surprised if your observed results did not match the probability. If you were to roll the die a very large number of times, you would expect that, overall, 26 of the rolls would result in an outcome of "at least five". You would not expect exactly 26. The long-term relative frequency of obtaining this result would approach the theoretical probability of 26 as the number of repetitions grows larger and larger.

This important characteristic of probability experiments is known as the **law of large numbers** which states that as the number of repetitions of an experiment is increased, the relative frequency obtained in the experiment tends to become closer and closer to the theoretical probability. Even though the outcomes do not happen according to any set pattern or order, overall, the long-term observed relative frequency will approach the theoretical probability. (The word **empirical** is often used instead of the word observed.)

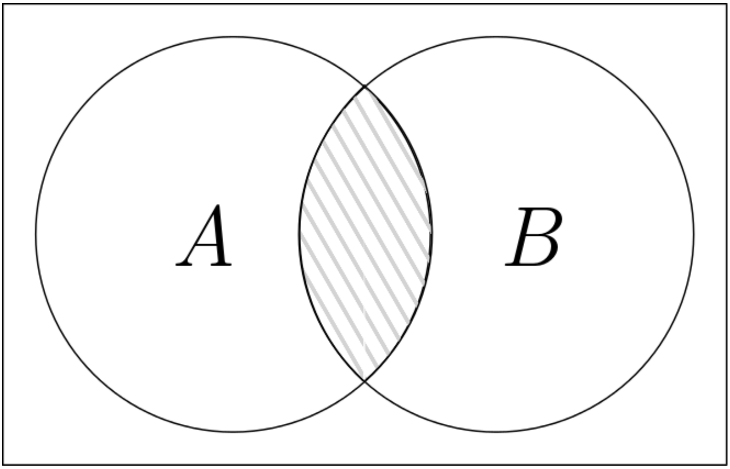
It is important to realize that in many situations, the outcomes are not equally likely. A coin or die may be **unfair**, or **biased**. Two math professors in Europe had their statistics students test the Belgian one Euro coin and discovered that in 250 trials, a head was obtained 56% of the time and a tail was obtained 44% of the time. The data seem to show that the coin is not a fair coin; more repetitions would be helpful to draw a more accurate conclusion about such bias. Some dice may be biased. Look at the dice in a game you have at home; the spots on each face are usually small holes carved out and then painted to make the spots visible. Your dice may or may not be biased; it is possible that the outcomes may be affected by the slight weight differences due to the different numbers of holes in the faces. Gambling casinos make a lot of money depending on outcomes from rolling dice, so casino dice are made differently to eliminate bias. Casino dice have flat faces; the holes are completely filled with paint having the same density as the material that the dice are made out of so that each face is equally likely to occur. Later we will learn techniques to use to work with probabilities for events that are not equally likely.

A key concept in probability is whether an event is **likely** or **unlikely**. A likely event is an event that has a good chance of happening, while an unlikely event is rare. For example, it is likely to snow in Calgary in the winter, but it is unlikely to snow in Calgary in the summer (it can happen, but it would be a rare or strange event). In general, in statistics, unlikely events usually have a probability of less than 1% of happening. Likely events usually have a probability of greater than 10% of happening. If the probability of the event is between 1% and 10%, it is up to the statistician or researcher to make a call to determine whether it is likely or unlikely.

"OR" Event:
An outcome is in the event A OR B if the outcome is in A or is in B or is in both A and B. For example, let A = {1, 2, 3, 4, 5} and B = {4, 5, 6, 7, 8}. A OR B = {1, 2, 3, 4, 5, 6, 7, 8}. Notice that 4 and 5 are NOT listed twice.



"AND" Event:
An outcome is in the event A AND B if the outcome is in both A and B at the same time. For example, let A and B be {1, 2, 3, 4, 5} and {4, 5, 6, 7, 8}, respectively. Then A AND B = {4, 5}.



The **complement** of event A is denoted A' (read "A prime"). A' consists of all outcomes that are **NOT** in A. Notice that $P(A) + P(A') = 1$. For example, let $S = \{1, 2, 3, 4, 5, 6\}$ and let $A = \{1, 2, 3, 4\}$. Then, $A' = \{5, 6\}$. $P(A) = 46$, $P(A') = 26$, and $P(A) + P(A') = 4\ 6 + 2\ 6 = 1$

The **conditional probability** of A given B is written $P(A|B)$. $P(A|B)$ is the probability that event A will occur given that the event B has already occurred. **A conditional reduces the sample space**. We calculate the probability of A from the reduced sample space B. The formula to calculate $P(A|B)$ is $P(A|B) = \mathrm{P(AANDB)\ P(B)}$ where $P(B)$ is greater than zero.

For example, suppose we toss one fair, six-sided die. The sample space $S = \{1, 2, 3, 4, 5, 6\}$. Let $A =$ face is 2 or 3 and $B =$ face is even (2, 4, 6). To calculate $P(A|B)$, we count the number of outcomes

2 or 3 in the sample space $B = \{2, 4, 6\}$. Then we divide that by the number of outcomes B (rather than S).

We get the same result by using the formula. Remember that S has six outcomes.

$P(A|B) =$ P(AANDB) P(B) $=$ (the number of outcomes that are 2 or 3 and even in S) 6 (the number of outcomes that are even in S) 6 $=$ 1 6 3 6 $=$ 1 3

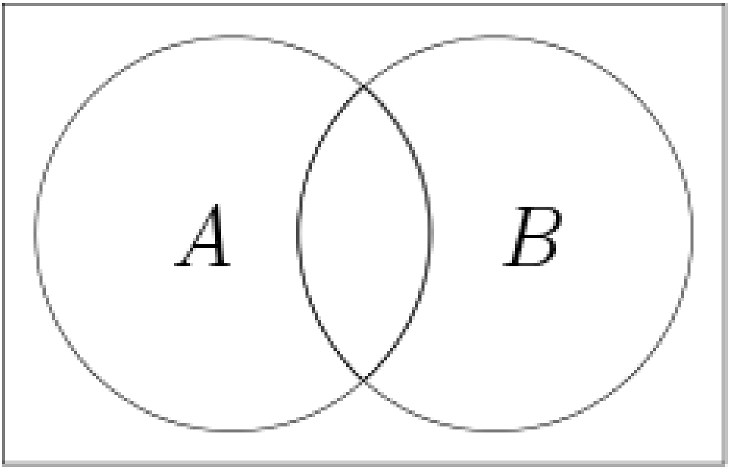
Odds
The odds of an event presents the probability as a ratio of success to failure. This is common in various gambling formats. Mathematically, the odds of an event can be defined as:
P(A) 1−P(A)

where $P(A)$ is the probability of success and of course $1 - P(A)$ is the probability of failure. Odds are always quoted as "numerator to denominator," e.g. 2 to 1. Here the probability of winning is twice that of losing; thus, the probability of winning is 0.66. A probability of winning of 0.60 would generate odds in favor of winning of 3 to 2. While the calculation of odds can be useful in gambling venues in determining payoff amounts, it is not helpful for understanding probability or statistical theory.

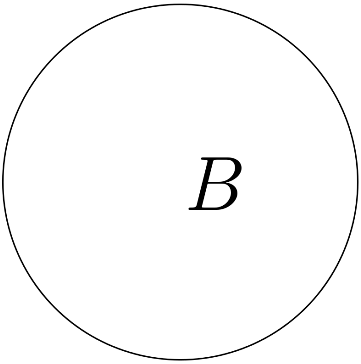
Understanding Terminology and Symbols

It is important to read each problem carefully to think about and understand what the events are. Understanding the wording is the first very important step in solving probability problems. Reread the problem several times if necessary. Clearly identify the event of interest. Determine whether there is a condition stated in the wording that would indicate that the probability is conditional; carefully identify the condition, if any.

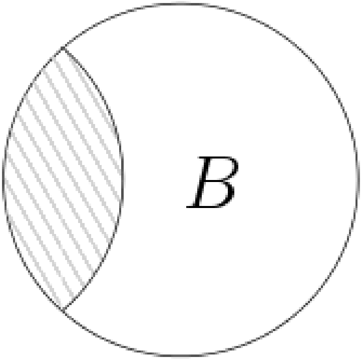
If the sample space is



then $P(A|B)$ is found by looking only at events that involved B:



and within *B* looking at the portion that involve *A*:



That portion is clearly the intersection of *A* and *B*.

The sample space *S* is the whole numbers starting at one and less than 20.

1. S = ____________________

   Let event A = the even numbers and event B = numbers greater than 13.
2. A = _______________, B = _______________
3. $P(A)$ = __________, $P(B)$ = ____________
4. A AND B = _______________, A OR B = ____________
5. $P(A$ AND $B)$ = _______, $P(A$ OR $B)$ = __________
6. A' = __________, $P(A')$ = __________
7. $P(A) + P(A')$ = _________
8. $P(A|B)$ = _________, $P(B|A)$ = _________; are the probabilities equal?

---

1. S = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19}
2. A = {2, 4, 6, 8, 10, 12, 14, 16, 18}, B = {14, 15, 16, 17, 18, 19}
3. $P(A)$ = 919 , $P(B)$ = 6 19
4. A AND B = {14,16,18}, A OR B = {2, 4, 6, 8, 10, 12, 14, 15, 16, 17, 18, 19}
5. $P(A$ AND $B)$ = 319 , $P(A$ OR $B)$ = 12 19
6. A' = 1, 3, 5, 7, 9, 11, 13, 15, 17, 19; $P(A')$ = 1019
7. $P(A) + P(A')$ = 1 (919 + 1019 = 1)
8. $P(A|B)$ = P(AANDB) P(B) = 3 6 , $P(B|A)$ = P(AANDB) P(A) = 3 9 , No

Try It

The sample space S is the ordered pairs of two whole numbers, the first from one to three and the second from one to four (Example: (1, 4)).

1. $S =$ ____________________

   Let event $A =$ the sum is even and event $B =$ the first number is prime.
2. $A =$ ______________, $B =$ ______________
3. $P(A) =$ __________, $P(B) =$ __________
4. A AND $B =$ ______________, A OR $B =$ ___________
5. $P(A \text{ AND } B) =$ ______, $P(A \text{ OR } B) =$ _________
6. $B' =$ _________, $P(B') =$ _________
7. $P(A) + P(A') =$ ________
8. $P(A|B) =$ ________, $P(B|A) =$ _________; are the probabilities equal?

1. $S = \{(1,1), (1,2), (1,3), (1,4), (2,1), (2,2), (2,3), (2,4), (3,1), (3,2), (3,3), (3,4)\}$
2. $A = \{(1,1), (1,3), (2,2), (2,4), (3,1), (3,3)\}$

   $B = \{(2,1), (2,2), (2,3), (2,4), (3,1), (3,2), (3,3), (3,4)\}$

3. $P(A) = 1\ 2$, $P(B) = 2\ 3$
4. A AND B = {(2,2), (2,4), (3,1), (3,3)}

   A OR B = {(1,1), (1,3), (2,1), (2,2), (2,3), (2,4), (3,1), (3,2), (3,3), (3,4)}
5. $P(A$ AND $B) = 1\ 3$, $P(A$ OR $B) = 5\ 6$
6. $B' =$ {(1,1), (1,2), (1,3), (1,4)}, $P(B') =$ 13
7. $P(B) + P(B') = 1$
8. $P(A|B) =$ P(A AND B) P(B) $= 1\ 2$, $P(B|A) =$ P(A AND B) P(B) $= 2\ 3$, No.

A fair, six-sided die is rolled. Describe the sample space S, identify each of the following events with a subset of S and compute its probability (an outcome is the number of dots that show up).

1. Event $T =$ the outcome is two.
2. Event $A =$ the outcome is an even number.
3. Event $B =$ the outcome is less than four.
4. The complement of A.
5. A GIVEN B
6. B GIVEN A
7. A AND B

8. A OR B
9. A OR B'
10. Event N = the outcome is a prime number.
11. Event I = the outcome is seven.

---

1. $T = \{2\}$, $P(T) = 16$
2. $A = \{2, 4, 6\}$, $P(A) = 12$
3. $B = \{1, 2, 3\}$, $P(B) = 12$
4. $A' = \{1, 3, 5\}$, $P(A') = 12$
5. $A|B = \{2\}$, $P(A|B) = 13$
6. $B|A = \{2\}$, $P(B|A) = 13$
7. A AND $B = \{2\}$, $P(A \text{ AND } B) = 16$
8. A OR $B = \{1, 2, 3, 4, 6\}$, $P(A \text{ OR } B) = 56$
9. A OR $B' = \{2, 4, 5, 6\}$, $P(A \text{ OR } B') = 23$
10. $N = \{2, 3, 5\}$, $P(N) = 12$
11. A six-sided die does not have seven dots. $P(7) = 0$.

[link] describes the distribution of a random sample S of 100 individuals, organized by gender and whether they are right- or left-handed.

	Right-handed	Left-handed
Males	43	9
Females	44	4

Let's denote the events $M =$ the subject is male, $F =$ the subject is female, $R =$ the subject is right-handed, $L =$ the subject is left-handed. Compute the following probabilities:

1. $P(M)$
2. $P(F)$
3. $P(R)$
4. $P(L)$
5. $P(M \text{ AND } R)$
6. $P(F \text{ AND } L)$
7. $P(M \text{ OR } F)$
8. $P(M \text{ OR } R)$
9. $P(F \text{ OR } L)$
10. $P(M')$
11. $P(R|M)$
12. $P(F|L)$
13. $P(L|F)$

1. $P(M) = 0.52$
2. $P(F) = 0.48$
3. $P(R) = 0.87$
4. $P(L) = 0.13$
5. $P(M \text{ AND } R) = 0.43$
6. $P(F \text{ AND } L) = 0.04$

7. $P(M \text{ OR } F) = 1$
8. $P(M \text{ OR } R) = 0.96$
9. $P(F \text{ OR } L) = 0.57$
10. $P(M') = 0.48$
11. $P(R|M) = 0.8269$ (rounded to four decimal places)
12. $P(F|L) = 0.3077$ (rounded to four decimal places)
13. $P(L|F) = 0.0833$

References

"Countries List by Continent." Worldatlas, 2013. Available online at http://www.worldatlas.com/cntycont.htm (accessed May 2, 2013).

Chapter Review

In this module we learned the basic terminology of probability. The set of all possible outcomes of an experiment is called the sample space. Events are subsets of the sample space, and they are assigned a probability that is a number between zero and one, inclusive.

Formula Review

A and B are events

$P(S) = 1$ where S is the sample space

$0 \leq P(A) \leq 1$

$P(A|B) =$ P(A ∩B) P(B)

In a particular college class, there are male and female students. Some students have long hair and some students have short hair. Write the **symbols** for the probabilities of the events for parts a through j. (Note that you cannot find numerical answers here. You were not given enough information to find any probability values yet; concentrate on understanding the symbols.)

- Let F be the event that a student is female.
- Let M be the event that a student is male.
- Let S be the event that a student has short hair.
- Let L be the event that a student has long hair.

1. The probability that a student does not

have long hair.

2. The probability that a student is male or has short hair.
3. The probability that a student is a female and has long hair.
4. The probability that a student is male, given that the student has long hair.
5. The probability that a student has long hair, given that the student is male.
6. Of all the female students, the probability that a student has short hair.
7. Of all students with long hair, the probability that a student is female.
8. The probability that a student is female or has long hair.
9. The probability that a randomly selected student is a male student with short hair.
10. The probability that a student is female.

---

1. $P(L') = P(S)$
2. $P(M \text{ OR } S)$
3. $P(F \text{ AND } L)$
4. $P(M|L)$
5. $P(L|M)$
6. $P(S|F)$
7. $P(F|L)$
8. $P(F \text{ OR } L)$
9. $P(M \text{ AND } S)$
10. $P(F)$

Use the following information to answer the next four exercises. A box is filled with several party favors. It contains 12 hats, 15 noisemakers, ten finger traps, and five bags of confetti.
Let $H =$ the event of getting a hat.
Let $N =$ the event of getting a noisemaker.
Let $F =$ the event of getting a finger trap.
Let $C =$ the event of getting a bag of confetti.

Find $P(H)$.

Find $P(N)$.

$P(N) = 15\ 42 = 5\ 14 = 0.36$

Find $P(F)$.

Find $P(C)$.

$P(C) = 5\ 42 = 0.12$

Use the following information to answer the next six exercises. A jar of 150 jelly beans contains 22 red jelly beans, 38 yellow, 20 green, 28 purple, 26 blue, and the rest are orange.

Let B = the event of getting a blue jelly bean
Let G = the event of getting a green jelly bean.
Let O = the event of getting an orange jelly bean.
Let P = the event of getting a purple jelly bean.
Let R = the event of getting a red jelly bean.
Let Y = the event of getting a yellow jelly bean.

Find $P(B)$.

Find $P(G)$.

$P(G) = 20\ 150 = 2\ 15 = 0.13$

Find $P(P)$.

Find $P(R)$.

$P(R) = 22\ 150 = 11\ 75 = 0.15$

Find $P(Y)$.

Find $P(O)$.

$P(O) = 150\text{-}22\text{-}38\text{-}20\text{-}28\text{-}26\ 150 = 16\ 150 = 8$

75 = 0.11

Use the following information to answer the next six exercises. There are 23 countries in North America, 12 countries in South America, 47 countries in Europe, 44 countries in Asia, 54 countries in Africa, and 14 in Oceania (Pacific Ocean region).
Let A = the event that a country is in Asia.
Let E = the event that a country is in Europe.
Let F = the event that a country is in Africa.
Let N = the event that a country is in North America.
Let O = the event that a country is in Oceania.
Let S = the event that a country is in South America.

Find $P(A)$.

Find $P(E)$.

$P(E)$ = 47 194 = 0.24

Find $P(F)$.

Find $P(N)$.

$P(N)$ = 23 194 = 0.12

Find $P(O)$.

Find $P(S)$.

$P(S)$ = 12 194 = 6 97 = 0.06

What is the probability of drawing a red card in a standard deck of 52 cards?

What is the probability of drawing a club in a standard deck of 52 cards?

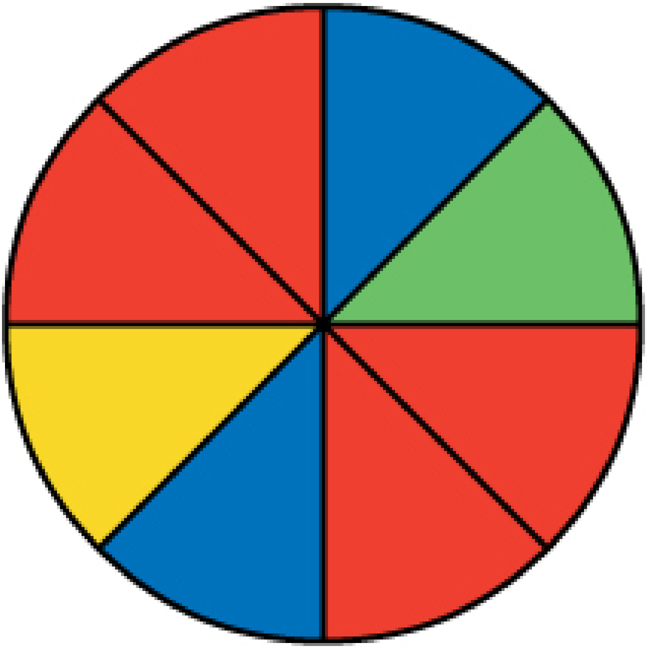
13 52 = 1 4 = 0.25

What is the probability of rolling an even number of dots with a fair, six-sided die numbered one through six?

What is the probability of rolling a prime number of dots with a fair, six-sided die numbered one through six?

3 6 = 1 2 = 0.5

Use the following information to answer the next two exercises. You see a game at a local fair. You have to throw a dart at a color wheel. Each section on the color wheel is equal in area.



Let B = the event of landing on blue.
Let R = the event of landing on red.
Let G = the event of landing on green.
Let Y = the event of landing on yellow.

If you land on Y, you get the biggest prize. Find $P(Y)$.

If you land on red, you don’t get a prize. What is *P*(*R*)?

P(R)= 4 8 =0.5

Use the following information to answer the next ten exercises. On a baseball team, there are infielders and outfielders. Some players are great hitters, and some players are not great hitters.
Let I = the event that a player in an infielder.
Let O = the event that a player is an outfielder.
Let H = the event that a player is a great hitter.
Let N = the event that a player is not a great hitter.

Write the symbols for the probability that a player is not an outfielder.

Write the symbols for the probability that a player is an outfielder or is a great hitter.

$P(O \text{ OR } H)$

Write the symbols for the probability that a player is an infielder and is not a great hitter.

Write the symbols for the probability that a player is a great hitter, given that the player is an infielder.

$P(H|I)$

Write the symbols for the probability that a player is an infielder, given that the player is a great hitter.

Write the symbols for the probability that of all the outfielders, a player is not a great hitter.

$P(N|O)$

Write the symbols for the probability that of all the great hitters, a player is an outfielder.

Write the symbols for the probability that a player is an infielder or is not a great hitter.

$P(I \text{ OR } N)$

Write the symbols for the probability that a player is an outfielder and is a great hitter.

Write the symbols for the probability that a player is an infielder.

$P(I)$

What is the word for the set of all possible outcomes?

What is conditional probability?

The likelihood that an event will occur given that another event has already occurred.

A shelf holds 12 books. Eight are fiction and the rest are nonfiction. Each is a different book with a unique title. The fiction books are numbered one to eight. The nonfiction books are numbered one to four. Randomly select one book
Let $F =$ event that book is fiction
Let $N =$ event that book is nonfiction
What is the sample space?

What is the sum of the probabilities of an event and its complement?

1

Use the following information to answer the next two exercises. You are rolling a fair, six-sided number cube. Let E = the event that it lands on an even number. Let M = the event that it lands on a multiple of three.

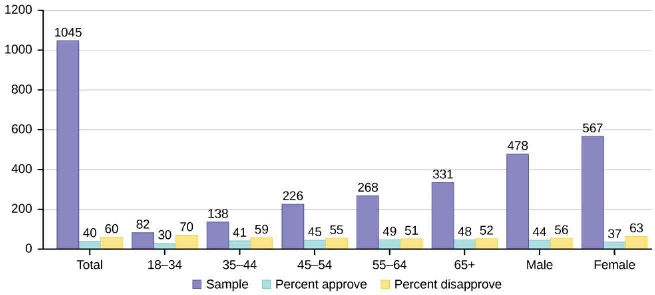
What does $P(E|M)$ mean in words?

What does $P(E \text{ OR } M)$ mean in words?

---

the probability of landing on an even number or a multiple of three

Homework



The graph in [link] displays the sample sizes and percentages of people in different age and gender groups who were polled concerning their approval of Mayor Ford's actions in office. The total number in the sample of all the age groups is 1,045.

1. Define three events in the graph.
2. Describe in words what the entry 40 means.
3. Describe in words the complement of the entry in question 2.
4. Describe in words what the entry 30 means.
5. Out of the males and females, what percent are males?
6. Out of the females, what percent disapprove of Mayor Ford?
7. Out of all the age groups, what percent approve of Mayor Ford?
8. Find $P(\text{Approve}|\text{Male})$.
9. Out of the age groups, what percent are more than 44 years old?
10. Find $P(\text{Approve}|\text{Age} < 35)$.

Explain what is wrong with the following statements. Use complete sentences.

1. If there is a 60% chance of rain on Saturday and a 70% chance of rain on

Sunday, then there is a 130% chance of rain over the weekend.
2. The probability that a baseball player hits a home run is greater than the probability that he gets a successful hit.

---

1. You can't calculate the joint probability knowing the probability of both events occurring, which is not in the information given; the probabilities should be multiplied, not added; and probability is never greater than 100%
2. A home run by definition is a successful hit, so he has to have at least as many successful hits as home runs.

Glossary

Conditional Probability
: the likelihood that an event will occur given that another event has already occurred

Equally Likely
: Each outcome of an experiment has the same probability.

Event
: a subset of the set of all outcomes of an experiment; the set of all outcomes of an

experiment is called a **sample space** and is usually denoted by S. An event is an arbitrary subset in S. It can contain one outcome, two outcomes, no outcomes (empty subset), the entire sample space, and the like. Standard notations for events are capital letters such as A, B, C, and so on.

Experiment

a planned activity carried out under controlled conditions

Outcome

a particular result of an experiment

Probability

a number between zero and one, inclusive, that gives the likelihood that a specific event will occur; the foundation of statistics is given by the following 3 axioms (by A.N. Kolmogorov, 1930's): Let S denote the sample space and A and B are two events in S. Then:

- $0 \leq P(A) \leq 1$
- If A and B are any two mutually exclusive events, then $P(A \text{ OR } B) = P(A) + P(\text{B})$.
- $P(S) = 1$

Sample Space

the set of all possible outcomes of an experiment

The Intersection: the AND Event

An outcome is in the event A AND B if the outcome is in both A AND B at the same time.

The Complement Event

The complement of event A consists of all outcomes that are NOT in A.

The Conditional Probability of A GIVEN B

$P(A|B)$ is the probability that event A will occur given that the event B has already occurred.

The Union: the OR Event

An outcome is in the event A OR B if the outcome is in A or is in B or is in both A and B.

Independent and mutually exclusive do **not** mean the same thing.

Independent Events

Two events are independent if the following are true:

- $P(A|B) = P(A)$
- $P(B|A) = P(B)$
- $P(A \text{ AND } B) = P(A)P(B)$

Two events A and B are **independent** if the knowledge that one occurred does not affect the chance the other occurs. For example, the outcomes of two roles of a fair die are independent events. The outcome of the first roll does not change the probability for the outcome of the second roll. To show two events are independent, you must show **only one** of the above conditions. If two events are NOT independent, then we say that they are **dependent**.

Sampling may be done **with replacement** or **without replacement**.

- **With replacement**: If each member of a population is replaced after it is picked, then that member has the possibility of being chosen more than once. When sampling is done with replacement, then events are considered to be independent, meaning the result of the first pick will not change the probabilities for the second pick.

- **Without replacement**: When sampling is done without replacement, each member of a population may be chosen only once. In this case, the probabilities for the second pick are affected by the result of the first pick. The events are considered to be dependent or not independent.

If it is not known whether *A* and *B* are independent or dependent, **assume they are dependent until you can show otherwise**.

You have a fair, well-shuffled deck of 52 cards. It consists of four suits. The suits are clubs, diamonds, hearts and spades. There are 13 cards in each suit consisting of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, *J* (jack), *Q* (queen), *K* (king) of that suit.

a. Sampling with replacement:

Suppose you pick three cards with replacement. The first card you pick out of the 52 cards is the *Q* of spades. You put this card back, reshuffle the cards and pick a second card from the 52-card deck. It is the ten of clubs. You put this card back, reshuffle the cards and pick a third card from the 52-card deck. This time, the card is the *Q* of spades again. Your picks are {*Q* of spades, ten of clubs, *Q* of spades}. You have picked the *Q* of spades twice. You pick each card from the 52-card deck.

b. Sampling without replacement:

Suppose you pick three cards without replacement. The first card you pick out of the 52 cards is the *K* of hearts. You put this card aside and pick the second card from the 51 cards remaining in the deck. It is the three of diamonds. You put this card aside and pick the third card from the remaining 50 cards in the deck. The third card is the *J* of spades. Your picks are {*K* of hearts, three of diamonds, *J* of spades}. Because you have picked the cards without replacement, you cannot pick the same card twice. The probability of picking the three of diamonds is called a conditional probability because it is conditioned on what was picked first. This is true also of the probability of picking the J of spades. The probability of picking the J of spades is actually conditioned on *both* the previous picks.

Try It

You have a fair, well-shuffled deck of 52 cards. It consists of four suits. The suits are clubs, diamonds, hearts and spades. There are 13 cards in each suit consisting of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, *J* (jack), *Q* (queen), *K* (king) of that suit. Three cards are picked at random.

1. Suppose you know that the picked cards are *Q* of spades, *K* of hearts and *Q* of spades. Can you decide if the sampling was with or without replacement?
2. Suppose you know that the picked cards are *Q* of spades, *K* of hearts, and *J* of spades. Can you decide if the sampling was with or without replacement?

1. With replacement
2. No

You have a fair, well-shuffled deck of 52 cards. It consists of four suits. The suits are clubs, diamonds, hearts, and spades. There are 13 cards in each suit consisting of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, *J* (jack), *Q* (queen), and *K* (king) of that suit. $S =$ spades, $H =$ Hearts, $D =$ Diamonds, $C =$ Clubs.

1. Suppose you pick four cards, but do not

put any cards back into the deck. Your cards are *QS*, 1*D*, 1*C*, *QD*.

2. Suppose you pick four cards and put each card back before you pick the next card. Your cards are *KH*, 7*D*, 6*D*, *KH*.

Which of a. or b. did you sample with replacement and which did you sample without replacement?

a. Without replacement; b. With replacement

Try It

You have a fair, well-shuffled deck of 52 cards. It consists of four suits. The suits are clubs, diamonds, hearts, and spades. There are 13 cards in each suit consisting of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, *J* (jack), *Q* (queen), and *K* (king) of that suit. $S =$ spades, $H =$ Hearts, $D =$ Diamonds, $C =$ Clubs. Suppose that you sample four cards without replacement. Which of the following outcomes are possible? Answer the same question for sampling with replacement.

1. *QS*, 1*D*, 1*C*, *QD*

2. *KH*, 7*D*, 6*D*, *KH*
3. *QS*, 7*D*, 6*D*, *KS*

without replacement: 1. Possible; 2. Impossible, 3. Possible

with replacement: 1. Possible; 2. Possible, 3. Possible

Mutually Exclusive Events

A and *B* are **mutually exclusive** events if they cannot occur at the same time. This means that *A* and *B* do not share any outcomes and *P*(*A* AND *B*) = 0.

For example, suppose the sample space *S* = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}. Let *A* = {1, 2, 3, 4, 5}, *B* = {4, 5, 6, 7, 8}, and *C* = {7, 9}. *A* AND *B* = {4, 5}. *P*(*A* AND *B*) = 210 and is not equal to zero. Therefore, *A* and *B* are not mutually exclusive. *A* and *C* do not have any numbers in common so *P*(*A* AND *C*) = 0. Therefore, *A* and *C* are mutually exclusive.

If it is not known whether *A* and *B* are mutually exclusive, **assume they are not until you can**

show otherwise. The following examples illustrate these definitions and terms.

Flip two fair coins. Find the probabilities of the events.

1. Let $F =$ the event of getting at most one tail (zero or one tail).
2. Let $G =$ the event of getting two faces that are the same.
3. Let $H =$ the event of getting a head on the first flip followed by a head or tail on the second flip.
4. Are F and G mutually exclusive?
5. Let $J =$ the event of getting all tails. Are J and H mutually exclusive?

Look at the sample space in [link].

1. Zero (0) or one (1) tails occur when the outcomes HH, TH, HT show up. $P(F) = 3$ 4
2. Two faces are the same if HH or TT show up. $P(G) = 24$
3. A head on the first flip followed by a head or tail on the second flip occurs when HH or HT show up. $P(H) = 24$

4. *F* and *G* share *HH* so *P*(*F* AND *G*) is not equal to zero (0). *F* and *G* are not mutually exclusive.
5. Getting all tails occurs when tails shows up on both coins (*TT*). *H*'s outcomes are *HH* and *HT*.

J and *H* have nothing in common so *P*(*J* AND *H*) $=$ 0. *J* and *H* are mutually exclusive.

Try It

A box has two balls, one white and one red. We select one ball, put it back in the box, and select a second ball (sampling with replacement). Find the probability of the following events:

1. Let *F* $=$ the event of getting the white ball twice.
2. Let *G* $=$ the event of getting two balls of different colors.
3. Let *H* $=$ the event of getting white on the first pick.
4. Are *F* and *G* mutually exclusive?
5. Are *G* and *H* mutually exclusive?

1. $P(F) = 14$
2. $P(G) = 12$
3. $P(H) = 12$
4. Yes
5. No

Roll one fair, six-sided die. The sample space is {1, 2, 3, 4, 5, 6}. Let event A = a face is odd. Then A = {1, 3, 5}. Let event B = a face is even. Then B = {2, 4, 6}.

- Find the complement of A, A'. The complement of A, A', is B because A and B together make up the sample space. $P(A) + P(B) = P(A) + P(A') = 1$. Also, $P(A) = 36$ and $P(B) = 36$.
- Let event C = odd faces larger than two. Then C = {3, 5}. Let event D = all even faces smaller than five. Then D = {2, 4}. $P(C$ AND $D) = 0$ because you cannot have an odd and even face at the same time. Therefore, C and D are mutually exclusive events.
- Let event E = all faces less than five. E = {1, 2, 3, 4}.

Are C and E mutually exclusive events?

(Answer yes or no.) Why or why not?

No. $C = \{3, 5\}$ and $E = \{1, 2, 3, 4\}$. $P(C$ AND $E) = 16$. To be mutually exclusive, $P(C$ AND $E)$ must be zero.

- Find $P(C|A)$. This is a conditional probability. Recall that the event C is $\{3, 5\}$ and event A is $\{1, 3, 5\}$. To find $P(C|A)$, find the probability of C using the sample space A. You have reduced the sample space from the original sample space $\{1, 2, 3, 4, 5, 6\}$ to $\{1, 3, 5\}$. So, $P(C|A) = 2\ 3$.

Try It

Let event A = learning Spanish. Let event B = learning German. Then A AND B = learning Spanish and German. Suppose $P(A) = 0.4$ and $P(B) = 0.2$. $P(A$ AND $B) = 0.08$. Are events A and B independent? Hint: You must show ONE of the following:

- $P(A|B) = P(A)$
- $P(B|A)$
- $P(A$ AND $B) = P(A)P(B)$

$P(A|B) = \text{P(A AND B) P(B)} = 0.08\ 0.2$
$= 0.4 = \text{P(A)}$

The events are independent because $P(A|B) = P(A)$.

Let event $G =$ taking a math class. Let event $H =$ taking a science class. Then, G AND $H =$ taking a math class and a science class. Suppose $P(G) = 0.6$, $P(H) = 0.5$, and $P(G \text{ AND } H) = 0.3$. Are G and H independent?
If G and H are independent, then you must show **ONE** of the following:

- $P(G|H) = P(G)$
- $P(H|G) = P(H)$
- $P(G \text{ AND } H) = P(G)P(H)$

NOTE
The choice you make depends on the information you have. You could choose any of the methods here because you have the necessary information.

a. Show that $P(G|H) = P(G)$.

$P(G|H) = \text{P(G AND H) P(H)} = 0.3\ 0.5 = 0.6$
$= P(G)$

b. Show $P(G \text{ AND } H) = P(G)P(H)$.

$P(G)P(H) = (0.6)(0.5) = 0.3 = P(G \text{ AND } H)$

Since G and H are independent, knowing that a person is taking a science class does not change the chance that he or she is taking a math class. If the two events had not been independent (that is, they are dependent) then knowing that a person is taking a science class would change the chance he or she is taking math. For practice, show that $P(H|G) = P(H)$ to show that G and H are independent events.

Try It

In a bag, there are six red marbles and four green marbles. The red marbles are marked with the numbers 1, 2, 3, 4, 5, and 6. The green marbles are marked with the numbers 1, 2, 3, and 4.

- R = a red marble
- G = a green marble
- O = an odd-numbered marble
- The sample space is $S = \{R1, R2, R3, R4, R5, R6, G1, G2, G3, G4\}$.

S has ten outcomes. What is $P(G \text{ AND } O)$?

Event G and $O = \{G1, G3\}$

$P(G \text{ and } O) = 2\ 10 = 0.2$

Let event C = taking an English class. Let event D = taking a speech class.

Suppose $P(C) = 0.75$, $P(D) = 0.3$, $P(C|D) = 0.75$ and $P(C \text{ AND } D) = 0.225$.

Justify your answers to the following questions numerically.

1. Are C and D independent?
2. Are C and D mutually exclusive?
3. What is $P(D|C)$?

1. Yes, because $P(C|D) = P(C)$.

2. No, because $P(C \text{ AND } D)$ is not equal to zero.
3. $P(D|C) = \text{P(C AND D) P(C)} = 0.225\ 0.75 = 0.3$

Try It

A student goes to the library. Let events $B =$ the student checks out a book and $D =$ the student checks out a DVD. Suppose that $P(B) = 0.40$, $P(D) = 0.30$ and $P(B \text{ AND } D) = 0.20$.

1. Find $P(B|D)$.
2. Find $P(D|B)$.
3. Are B and D independent?
4. Are B and D mutually exclusive?

1. $P(B|D) = 0.6667$
2. $P(D|B) = 0.5$
3. No
4. No

In a box there are three red cards and five blue

cards. The red cards are marked with the numbers 1, 2, and 3, and the blue cards are marked with the numbers 1, 2, 3, 4, and 5. The cards are well-shuffled. You reach into the box (you cannot see into it) and draw one card.
Let R = red card is drawn, B = blue card is drawn, E = even-numbered card is drawn.
The sample space S = $R1$, $R2$, $R3$, $B1$, $B2$, $B3$, $B4$, $B5$. S has eight outcomes.

- $P(R)$ = 3 8 . $P(B)$ = 5 8 . $P(R$ AND $B)$ = 0. (You cannot draw one card that is both red and blue.)
- $P(E)$ = 3 8 . (There are three even-numbered cards, $R2$, $B2$, and $B4$.)
- $P(E|B)$ = 2 5 . (There are five blue cards: $B1$, $B2$, $B3$, $B4$, and $B5$. Out of the blue cards, there are two even cards; $B2$ and $B4$.)
- $P(B|E)$ = 2 3 . (There are three even-numbered cards: $R2$, $B2$, and $B4$. Out of the even-numbered cards, to are blue; $B2$ and $B4$.)
- The events R and B are mutually exclusive because $P(R$ AND $B)$ = 0.
- Let G = card with a number greater than 3. G = {$B4$, $B5$}. $P(G)$ = 2 8 . Let H = blue card numbered between one and four, inclusive. H = {$B1$, $B2$, $B3$, $B4$}. $P(G|H)$ = 1 4 . (The only card in H that has a number greater than three is $B4$.) Since 2 8 = 1 4 , $P(G) = P(G|H)$, which means that G and H are independent.

Try It

In a basketball arena,

- 70% of the fans are rooting for the home team.
- 25% of the fans are wearing blue.
- 20% of the fans are wearing blue and are rooting for the away team.
- Of the fans rooting for the away team, 67% are wearing blue.

Let A be the event that a fan is rooting for the away team.
Let B be the event that a fan is wearing blue.
Are the events of rooting for the away team and wearing blue independent? Are they mutually exclusive?

$P(B|A) = 0.67$

$P(B) = 0.25$

So $P(B)$ does not equal $P(B|A)$ which means that B and A are not independent (wearing blue and rooting for the away team are not independent). They are also not mutually exclusive, because $P(\text{B AND } A) = 0.20$, not 0.

In a particular college class, 60% of the students are female. Fifty percent of all students in the class have long hair. Forty-five percent of the students are female and have long hair. Of the female students, 75% have long hair. Let *F* be the event that a student is female. Let *L* be the event that a student has long hair. One student is picked randomly. Are the events of being female and having long hair independent?

- The following probabilities are given in this example:
- $P(F) = 0.60$; $P(L) = 0.50$
- $P(F \text{ AND } L) = 0.45$
- $P(L|F) = 0.75$

NOTE

The choice you make depends on the information you have. You could use the first or last condition on the list for this example. You do not know $P(F|L)$ yet, so you cannot use the second condition.

Solution 1

Check whether $P(F \text{ AND } L) = P(F)P(L)$. We are given that $P(F \text{ AND } L) = 0.45$, but $P(F)P(L) = (0.60)(0.50) = 0.30$. The events of being female and having long hair are not independent because $P(F \text{ AND } L)$ does not equal $P(F)P(L)$.

Solution 2

Check whether $P(L|F)$ equals $P(L)$. We are given that $P(L|F) = 0.75$, but $P(L) = 0.50$; they are not equal. The events of being female and having long hair are not independent.

Interpretation of Results

The events of being female and having long hair are not independent; knowing that a student is female changes the probability that a student has long hair.

Try It

Mark is deciding which route to take to work. His choices are $I =$ the Interstate and $F =$ Fifth Street.

- $P(I) = 0.44$ and $P(F) = 0.55$
- $P(I \text{ AND } F) = 0$ because Mark will take only one route to work.

What is the probability of $P(I \text{ OR } F)$?

Because $P(I \text{ AND } F) = 0$,

$P(I \text{ OR } F) = P(I) + P(F) - P(I \text{ AND } F) = 0.44$

$+ 0.56 - 0 = 1$

1. Toss one fair coin (the coin has two sides, *H* and *T*). The outcomes are ______. Count the outcomes. There are ___ outcomes.
2. Toss one fair, six-sided die (the die has 1, 2, 3, 4, 5 or 6 dots on a side). The outcomes are ____________. Count the outcomes. There are __ outcomes.
3. Multiply the two numbers of outcomes. The answer is _____.
4. If you flip one fair coin and follow it with the toss of one fair, six-sided die, the answer in three is the number of outcomes (size of the sample space). What are the outcomes? (Hint: Two of the outcomes are *H*1 and *T*6.)
5. Event A = heads (*H*) on the coin followed by an even number (2, 4, 6) on the die. $A = \{$_____________$\}$. Find $P(A)$.
6. Event B = heads on the coin followed by a three on the die. $B = \{$______$\}$. Find $P(B)$.
7. Are A and B mutually exclusive? (Hint: What is $P(A \text{ AND } B)$? If $P(A \text{ AND } B) = 0$, then A and B are mutually exclusive.)
8. Are A and B independent? (Hint: Is *P*(*A*

AND *B*) = *P*(*A*)*P*(*B*)? If *P*(*A* AND *B*) = *P*(*A*)*P*(*B*), then *A* and *B* are independent. If not, then they are dependent).

1. *H* and *T*; 2
2. 1, 2, 3, 4, 5, 6; 6
3. 2(6) = 12
4. *T*1, *T*2, *T*3, *T*4, *T*5, *T*6, *H*1, *H*2, *H*3, *H*4, *H*5, *H*6
5. *A* = {*H*2, *H*4, *H*6}; *P*(*A*) = 3 12
6. *B* = {*H*3}; *P*(*B*) = 1 12
7. Yes, because *P*(*A* AND *B*) = 0
8. *P*(*A* AND *B*) = 0.*P*(*A*)*P*(*B*) = (3 12) (1 12) . *P*(*A* AND *B*) does not equal *P*(*A*)*P*(*B*), so *A* and *B* are dependent.

Try It

A box has two balls, one white and one red. We select one ball, put it back in the box, and select a second ball (sampling with replacement). Let *T* be the event of getting the white ball twice, *F* the event of picking the white ball first, *S* the event of picking the white ball in the second drawing.

1. Compute $P(T)$.
2. Compute $P(T|F)$.
3. Are T and F independent?.
4. Are F and S mutually exclusive?
5. Are F and S independent?

---

1. $P(T) = 1\ 4$
2. $P(T|F) = 1\ 2$
3. No
4. No
5. Yes

References

Lopez, Shane, Preety Sidhu. “U.S. Teachers Love Their Lives, but Struggle in the Workplace.” Gallup Wellbeing, 2013. http://www.gallup.com/poll/161516/teachers-love-lives-struggle-workplace.aspx (accessed May 2, 2013).

Data from Gallup. Available online at www.gallup.com/ (accessed May 2, 2013).

Chapter Review

Two events A and B are independent if the knowledge that one occurred does not affect the chance the other occurs. If two events are not independent, then we say that they are dependent.

In sampling with replacement, each member of a population is replaced after it is picked, so that member has the possibility of being chosen more than once, and the events are considered to be independent. In sampling without replacement, each member of a population may be chosen only once, and the events are considered not to be independent. When events do not share outcomes, they are mutually exclusive of each other.

Formula Review

If A and B are independent, $P(A \cap B) = P(A)P(B)$, $P(A|B) = P(A)$ and $P(B|A) = P(B)$.

If A and B are mutually exclusive, $P(A \cup B) = P(A) + P(B)$ and $P(A \text{ AND } B) = 0$.

E and F are mutually exclusive events. $P(E) = 0.4$; $P(F) = 0.5$. Find $P(E|F)$.

J and K are independent events. $P(J|K) = 0.3$. Find $P(J)$.

$P(J) = 0.3$

U and V are mutually exclusive events. $P(U) = 0.26$; $P(V) = 0.37$. Find:

1. $P(U \text{ AND } V) =$
2. $P(U|V) =$
3. $P(U \text{ OR } V) =$

Q and R are independent events. $P(Q) = 0.4$ and $P(Q \text{ AND } R) = 0.1$. Find $P(R)$.

$P(Q \text{ AND } R) = P(Q)P(R)$

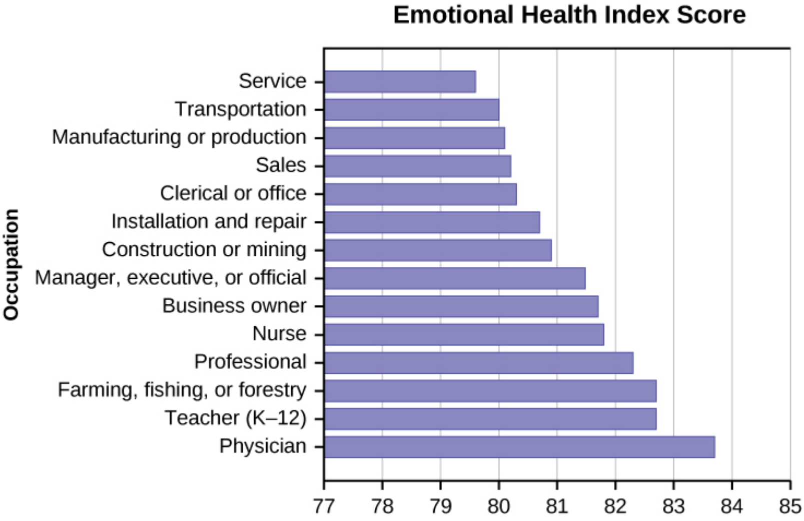
$0.1 = (0.4)P(R)$

$P(R) = 0.25$

Homework

Use the following information to answer the next 12 exercises. The graph shown is based on more than

170,000 interviews done by Gallup that took place from January through December 2012. The sample consists of employed Americans 18 years of age or older. The Emotional Health Index Scores are the sample space. We randomly sample one Emotional Health Index Score.



Find the probability that an Emotional Health Index Score is 82.7.

Find the probability that an Emotional Health Index Score is 81.0.

0

Find the probability that an Emotional Health Index Score is more than 81?

Find the probability that an Emotional Health Index Score is between 80.5 and 82?

0.3571

If we know an Emotional Health Index Score is 81.5 or more, what is the probability that it is 82.7?

What is the probability that an Emotional Health Index Score is 80.7 or 82.7?

0.2142

What is the probability that an Emotional Health Index Score is less than 80.2 given that it is already less than 81.

What occupation has the highest emotional index score?

Physician (83.7)

What occupation has the lowest emotional index score?

What is the range of the data?

$83.7 - 79.6 = 4.1$

Compute the average EHIS.

If all occupations are equally likely for a certain individual, what is the probability that he or she will have an occupation with lower than average EHIS?

$P(\text{Occupation} < 81.3) = 0.5$

Bringing It Together

A previous year, the weights of the members of the **San Francisco 49ers** and the **Dallas Cowboys** were published in the *San Jose Mercury News*. The factual data are compiled into [link].

Shirt#	≤ 210	211–250	251–290	290 ≤
1–33	21	5	0	0
34–66	6	18	7	4
66–99	6	12	22	5

For the following, suppose that you randomly select one player from the 49ers or Cowboys.

If having a shirt number from one to 33 and weighing at most 210 pounds were independent events, then what should be true about $P(\text{Shirt\# 1–33}|\leq 210 \text{ pounds})$?

The probability that a male develops some form of cancer in his lifetime is 0.4567. The probability that a male has at least one false positive test result (meaning the test comes back for cancer when the man does not have it) is 0.51. Some of the following questions do not have enough information for you to answer them. Write “not enough information” for those answers. Let C = a man develops cancer in his lifetime and P = man has at least one false positive.

1. $P(C)$ = ____
2. $P(P|C)$ = ____
3. $P(P|C')$ = ____
4. If a test comes up positive, based upon numerical values, can you assume that

man has cancer? Justify numerically and explain why or why not.

---

1. $P(C) = 0.4567$
2. not enough information
3. not enough information
4. No, because over half (0.51) of men have at least one false positive text

Given events G and H: $P(G) = 0.43$; $P(H) = 0.26$; $P(H \text{ AND } G) = 0.14$

1. Find $P(H \text{ OR } G)$.
2. Find the probability of the complement of event (H AND G).
3. Find the probability of the complement of event (H OR G).

Given events J and K: $P(J) = 0.18$; $P(K) = 0.37$; $P(J \text{ OR } K) = 0.45$

1. Find $P(J \text{ AND } K)$.
2. Find the probability of the complement of event (J AND K).
3. Find the probability of the complement of event (J AND K).

---

1. $P(J \text{ OR } K) = P(J) + P(K) - P(J \text{ AND } K)$; $0.45 = 0.18 + 0.37$ - $P(J \text{ AND } K)$; solve to find $P(J \text{ AND } K) = 0.10$
2. $P(\text{NOT } (J \text{ AND } K)) = 1$ - $P(J \text{ AND } K) = 1$ - $0.10 = 0.90$
3. $P(\text{NOT } (J \text{ OR } K)) = 1$ - $P(J \text{ OR } K) = 1$ - $0.45 = 0.55$

Glossary

Dependent Events

If two events are NOT independent, then we say that they are dependent.

Sampling with Replacement

If each member of a population is replaced after it is picked, then that member has the possibility of being chosen more than once.

Sampling without Replacement

When sampling is done without replacement, each member of a population may be chosen only once.

Two Basic Rules of Probability

When calculating probability, there are two rules to consider when determining if two events are independent or dependent and if they are mutually exclusive or not.

The Multiplication Rule

If *A* and *B* are two events defined on a **sample space**, then: $P(A \cap B) = P(B)P(A \mid B)$. We can think of the intersection symbol as substituting for the word "and".

This rule may also be written as: $P(A \mid B) = P(A \cap B)$
$P(B)$

This equation is read as the probability of *A* given *B* equals the probability of *A* and *B* divided by the probability of *B*.

If *A* and *B* are **independent**, then $P(A \mid B) = P(A)$. Then $P(A \cap B) = P(A \mid B)P(B)$ becomes $P(A \cap B) = P(A)(B)$ because the $P(A \mid B) = P(A)$ if *A* and *B* are independent.

One easy way to remember the multiplication rule is that the word "and" means that the event has to satisfy two conditions. For example the name drawn from the class roster is to be both a female and a sophomore. It is harder to satisfy two conditions than only one and of course when we multiply fractions the result is always smaller. This reflects the increasing difficulty of satisfying two conditions.

The Addition Rule

If *A* and *B* are defined on a sample space, then: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. We can think of the union symbol substituting for the word "or". The reason we subtract the intersection of *A* and *B* is to keep from double counting elements that are in both *A* and *B*.

If *A* and *B* are **mutually exclusive**, then $P(A \cap B) = 0$. Then $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ becomes $P(A \cup B) = P(A) + P(B)$.

Klaus is trying to choose where to go on vacation. His two choices are: A = New Zealand and B = Alaska

- Klaus can only afford one vacation. The probability that he chooses *A* is $P(A) = 0.6$ and the probability that he chooses *B* is $P(B) = 0.35$.
- $P(A \cap B) = 0$ because Klaus can only afford

to take one vacation

- Therefore, the probability that he chooses either New Zealand or Alaska is P (A ∪ B) = P (A) + P (B) = 0.6 + 0.35 = 0.95. Note that the probability that he does not choose to go anywhere on vacation must be 0.05.

Carlos plays college soccer. He makes a goal 65% of the time he shoots. Carlos is going to attempt two goals in a row in the next game. A = the event Carlos is successful on his first attempt. $P(A) = 0.65$. B = the event Carlos is successful on his second attempt. $P(B) = 0.65$. Carlos tends to shoot in streaks. The probability that he makes the second goal | that he made the first goal is 0.90.

a. What is the probability that he makes both goals?

a. The problem is asking you to find P (A ∩ B) = P (B ∩ A) . Since $P(B|A) = 0.90$: $P(B \cap A) = P(B|A)\,P(A) = (0.90)(0.65) = 0.585$

Carlos makes the first and second goals with probability 0.585.

b. What is the probability that Carlos makes either the first goal or the second goal?

b. The problem is asking you to find $P(A \cup B)$.

$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.65 + 0.65 - 0.585 = 0.715$

Carlos makes either the first goal or the second goal with probability 0.715.

c. Are A and B independent?

c. No, they are not, because $P(B \cap A) = 0.585$.

$P(B)P(A) = (0.65)(0.65) = 0.423$

$0.423 \neq 0.585 = P(B \cap A)$

So, $P(B \cap A)$ is **not** equal to $P(B)P(A)$.

d. Are A and B mutually exclusive?

d. No, they are not because $P(A \cap B) = 0.585$.

To be mutually exclusive, $P(A \cap B)$ must equal zero.

Try It

Helen plays basketball. For free throws, she makes the shot 75% of the time. Helen must now attempt two free throws. C = the event that Helen makes the first shot. $P(C) = 0.75$. D = the event Helen makes the second shot. $P(D) = 0.75$. The probability that Helen makes the second free throw given that she made the first is 0.85. What is the probability that Helen makes both free throws?

$P(D|C) = 0.85$

$P(C \cap D) = P(D \cap C)$
$P(D \cap C) = P(D|C)P(C) = (0.85)(0.75) = 0.6375$
Helen makes the first and second free throws with probability 0.6375.

A community swim team has **150** members. **Seventy-five** of the members are advanced swimmers. **Forty-seven** of the members are intermediate swimmers. The remainder are novice swimmers. **Forty** of the advanced swimmers practice four times a week. **Thirty** of the intermediate swimmers practice four times a week. **Ten** of the novice swimmers practice four times a week. Suppose one member of the swim team is chosen randomly.

a. What is the probability that the member is a novice swimmer?

a. 28150

b. What is the probability that the member practices four times a week?

b. 80150

c. What is the probability that the member is an advanced swimmer and practices four times a week?

c. 40150

d. What is the probability that a member is an advanced swimmer and an intermediate swimmer? Are being an advanced swimmer and an intermediate swimmer mutually exclusive? Why or why not?

d. $P(\text{advanced} \cap \text{intermediate}) = 0$, so these are mutually exclusive events. A swimmer cannot be an advanced swimmer and an intermediate swimmer at the same time.

e. Are being a novice swimmer and practicing four times a week independent events? Why or why not?

e. No, these are not independent events.
$P(\text{novice} \cap \text{practices four times per week}) = 0.0667$
$P(\text{novice})P(\text{practices four times per week}) = 0.0996$
$0.0667 \neq 0.0996$

Try It

A school has 200 seniors of whom 140 will be going to college next year. Forty will be going directly to work. The remainder are taking a gap year. Fifty of the seniors going to college play sports. Thirty of the seniors going directly to work play sports. Five of the seniors taking a gap year play sports. What is the probability that a senior is taking a gap year?

$P= 200-140-40\ 200 = 20\ 200 = 0.1$

Felicity attends Modesto JC in Modesto, CA. The probability that Felicity enrolls in a math class is 0.2 and the probability that she enrolls in a speech class is 0.65. The probability that she enrolls in a math class | that she enrolls in speech class is 0.25. Let: M = math class, S = speech class, $M|S$ = math given speech

1. What is the probability that Felicity enrolls in math and speech?
   Find $P(M \cap S) = P(M|S)P(S)$.
2. What is the probability that Felicity enrolls in math or speech classes?
   Find $P(M \cup S) = P(M) + P(S)$ - $P(M \cap S)$.

3. Are M and S independent? Is $P(M|S) = P(M)$?
4. Are M and S mutually exclusive? Is $P(M \cap S) = 0$?

a. 0.1625, b. 0.6875, c. No, d. No

Try It

A student goes to the library. Let events $B =$ the student checks out a book and $D =$ the student check out a DVD. Suppose that $P(B) = 0.40$, $P(D) = 0.30$ and $P(D|B) = 0.5$.

1. Find $P(B \cap D)$.
2. Find $P(B \cup D)$.

1. $P(B \cap D) = P(D|B)P(B) = (0.5)(0.4) = 0.20$.
2. $P(B \cup D) = P(B) + P(D) - P(B \cap D) = 0.40 + 0.30 - 0.20 = 0.50$

Studies show that about one woman in seven (approximately 14.3%) who live to be 90 will develop breast cancer. Suppose that of those women who develop breast cancer, a test is negative 2% of the time. Also suppose that in the general population of women, the test for breast cancer is negative about 85% of the time. Let $B =$ woman develops breast cancer and let $N =$ tests negative. Suppose one woman is selected at random.

a. What is the probability that the woman develops breast cancer? What is the probability that woman tests negative?

a. $P(B) = 0.143$; $P(N) = 0.85$

b. Given that the woman has breast cancer, what is the probability that she tests negative?

b. $P(N|B) = 0.02$

c. What is the probability that the woman has breast cancer AND tests negative?

c. $P(B \cap N) = P(B)P(N|B) = (0.143)(0.02) = 0.0029$

d. What is the probability that the woman has breast cancer or tests negative?

d. $P(B \cup N) = P(B) + P(N) - P(B \cap N) = 0.143 + 0.85 - 0.0029 = 0.9901$

e. Are having breast cancer and testing negative independent events?

e. No. $P(N) = 0.85$; $P(N|B) = 0.02$. So, $P(N|B)$ does not equal $P(N)$.

f. Are having breast cancer and testing negative mutually exclusive?

f. No. $P(B \cap N) = 0.0029$. For B and N to be mutually exclusive, $P(B \cap N)$ must be zero.

Try It

A school has 200 seniors of whom 140 will be going to college next year. Forty will be going directly to work. The remainder are taking a gap year. Fifty of the seniors going to college play sports. Thirty of the seniors going directly to work play sports. Five of the seniors taking a gap year play sports. What is the probability that a senior is going to college and plays sports?

Let A = student is a senior going to college.

Let B = student plays sports.

$P(B) = 140 \ 200$

$P(B|A) = 50 \ 140$

$P(A \cap B) = P(B|A)P(A)$

$P(A \cap B) = (\ 140 \ 200 \ )(\ 50 \ 140 \ ) = 1 \ 4$

Refer to the information in [link]. P = tests positive.

1. Given that a woman develops breast cancer, what is the probability that she tests positive. Find $P(P|B) = 1 - P(N|B)$.
2. What is the probability that a woman develops breast cancer and tests positive. Find $P(B \cap P) = P(P|B)P(B)$.
3. What is the probability that a woman does not develop breast cancer. Find $P(B') = 1 - P(B)$.
4. What is the probability that a woman tests positive for breast cancer. Find $P(P) = 1 - P(N)$.

a. 0.98; b. 0.1401; c. 0.857; d. 0.15

Try It

A student goes to the library. Let events B = the student checks out a book and D = the student checks out a DVD. Suppose that $P(B) = 0.40$, $P(D) = 0.30$ and $P(D|B) = 0.5$.

1. Find $P(B')$.
2. Find $P(D \cap B)$.
3. Find $P(B|D)$.
4. Find $P(D \cap B')$.
5. Find $P(D|B')$.

1. $P(B') = 0.60$
2. $P(D \cap B) = P(D|B)P(B) = 0.20$
3. $P(B|D) =$ P(B∩D) P(D) $=$ (0.20) (0.30) $= 0.66$
4. $P(D \cap B') = P(D) - P(D \cap B) = 0.30 - 0.20 = 0.10$
5. $P(D|B') = P(D \cap B')P(B') = (P(D) - P(D \cap B))(0.60) = (0.10)(0.60) = 0.06$

References

DiCamillo, Mark, Mervin Field. “The File Poll.” Field Research Corporation. Available online at http://www.field.com/fieldpollonline/subscribers/Rls2443.pdf (accessed May 2, 2013).

Rider, David, “Ford support plummeting, poll suggests,” The Star, September 14, 2011. Available online at http://www.thestar.com/news/gta/2011/09/14/ford_support_plummeting_poll_suggests.html (accessed May 2, 2013).

“Mayor’s Approval Down.” News Release by Forum Research Inc. Available online at http://

www.forumresearch.com/forms/News Archives/News Releases/74209_TO_Issues_-_Mayoral_Approval_%28Forum_Research%29%2820130320%29.pdf (accessed May 2, 2013).

“Roulette.” Wikipedia. Available online at http://en.wikipedia.org/wiki/Roulette (accessed May 2, 2013).

Shin, Hyon B., Robert A. Kominski. “Language Use in the United States: 2007.” United States Census Bureau. Available online at http://www.census.gov/hhes/socdemo/language/data/acs/ACS-12.pdf (accessed May 2, 2013).

Data from the Baseball-Almanac, 2013. Available online at www.baseball-almanac.com (accessed May 2, 2013).

Data from U.S. Census Bureau.

Data from the Wall Street Journal.

Data from The Roper Center: Public Opinion Archives at the University of Connecticut. Available online at http://www.ropercenter.uconn.edu/ (accessed May 2, 2013).

Data from Field Research Corporation. Available online at www.field.com/fieldpollonline (accessed May 2,2 013).

Chapter Review

The multiplication rule and the addition rule are used for computing the probability of *A* and *B*, as well as the probability of *A* or *B* for two given events *A*, *B* defined on the sample space. In sampling with replacement each member of a population is replaced after it is picked, so that member has the possibility of being chosen more than once, and the events are considered to be independent. In sampling without replacement, each member of a population may be chosen only once, and the events are considered to be not independent. The events *A* and *B* are mutually exclusive events when they do not have any outcomes in common.

Formula Review

The multiplication rule: $P(A \cap B) = P(A|B)P(B)$

The addition rule: $P(A \cup B) = P(A) + P(B)$ - $P(A \cap B)$

Use the following information to answer the next ten exercises. Forty-eight percent of all Californians registered voters prefer life in prison without parole over the death penalty for a person convicted of first degree murder. Among Latino California registered voters, 55% prefer life in prison without parole over the death penalty for a person convicted of first degree murder. 37.6% of all Californians are Latino.

In this problem, let:

- C = Californians (registered voters) preferring life in prison without parole over the death penalty for a person convicted of first degree murder.
- L = Latino Californians

Suppose that one Californian is randomly selected.

Find $P(C)$.

Find $P(L)$.

0.376

Find $P(C|L)$.

In words, what is $C|L$?

$C|L$ means, given the person chosen is a Latino Californian, the person is a registered voter who prefers life in prison without parole for a person convicted of first degree murder.

Find $P(L \cap C)$.

In words, what is $L \cap C$?

$L \cap C$ is the event that the person chosen is a Latino California registered voter who prefers life without parole over the death penalty for a person convicted of first degree murder.

Are L and C independent events? Show why or why not.

Find $P(L \cup C)$.

0.6492

In words, what is $L \cup C$?

Are L and C mutually exclusive events? Show why or why not.

No, because $P(L \cap C)$ does not equal 0.

(credit: film8ker/wikibooks)

Homework

On February 28, 2013, a Field Poll Survey reported that 61% of California registered voters approved of allowing two people of the same gender to marry and have regular marriage laws apply to them. Among 18 to 39 year olds (California registered voters), the approval rating was 78%. Six in ten California registered voters said that the upcoming Supreme Court's ruling about the constitutionality of California's Proposition 8 was either very or somewhat important to them. Out of those CA registered voters who support same-sex marriage, 75% say the ruling is important to them.

In this problem, let:

- $C =$ California registered voters who support same-sex marriage.
- $B =$ California registered voters who say the Supreme Court's ruling about the constitutionality of California's Proposition 8 is very or somewhat important to them
- $A =$ California registered voters who are 18 to 39 years old.

1. Find $P(C)$.
2. Find $P(B)$.
3. Find $P(C|A)$.
4. Find $P(B|C)$.
5. In words, what is $C|A$?
6. In words, what is $B|C$?
7. Find $P(C \cap B)$.
8. In words, what is $C \cap B$?
9. Find $P(C \cup B)$.
10. Are C and B mutually exclusive events? Show why or why not.

After Rob Ford, the mayor of Toronto, announced his plans to cut budget costs in late 2011, the Forum Research polled 1,046 people to measure the mayor's popularity. Everyone polled expressed either approval or disapproval. These are the results their poll produced:

- In early 2011, 60 percent of the population approved of Mayor Ford's actions in office.
- In mid-2011, 57 percent of the population approved of his actions.
- In late 2011, the percentage of popular approval was measured at 42 percent.

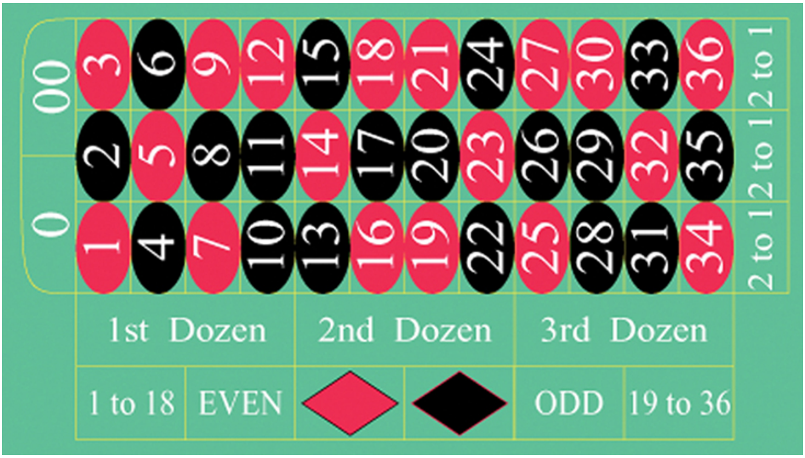
1. What is the sample size for this study?
2. What proportion in the poll disapproved of Mayor Ford, according to the results from late 2011?

3. How many people polled responded that they approved of Mayor Ford in late 2011?
4. What is the probability that a person supported Mayor Ford, based on the data collected in mid-2011?
5. What is the probability that a person supported Mayor Ford, based on the data collected in early 2011?

---

1. The Forum Research surveyed 1,046 Torontonians.
2. 58%
3. 42% of 1,046 = 439 (rounding to the nearest integer)
4. 0.57
5. 0.60.

Use the following information to answer the next three exercises. The casino game, roulette, allows the gambler to bet on the probability of a ball, which spins in the roulette wheel, landing on a particular color, number, or range of numbers. The table used to place bets contains of 38 numbers, and each number is assigned to a color and a range.



1. List the sample space of the 38 possible outcomes in roulette.
2. You bet on red. Find *P*(red).
3. You bet on -1st 12- (1st Dozen). Find *P*(-1st 12-).
4. You bet on an even number. Find *P*(even number).
5. Is getting an odd number the complement of getting an even number? Why?
6. Find two mutually exclusive events.
7. Are the events Even and 1st Dozen independent?

Compute the probability of winning the following types of bets:

1. Betting on two lines that touch each other on the table as in 1-2-3-4-5-6

2. Betting on three numbers in a line, as in 1-2-3
3. Betting on one number
4. Betting on four numbers that touch each other to form a square, as in 10-11-13-14
5. Betting on two numbers that touch each other on the table, as in 10-11 or 10-13
6. Betting on 0-00-1-2-3
7. Betting on 0-1-2; or 0-00-2; or 00-2-3

---

1. *P*(Betting on two line that touch each other on the table) = 6 38
2. *P*(Betting on three numbers in a line) = 3 38
3. *P*(Bettting on one number) = 1 38
4. *P*(Betting on four number that touch each other to form a square) = 4 38
5. *P*(Betting on two number that touch each other on the table) = 2 38
6. *P*(Betting on 0-00-1-2-3) = 5 38
7. *P*(Betting on 0-1-2; or 0-00-2; or 00-2-3) = 3 38

Compute the probability of winning the following types of bets:

1. Betting on a color
2. Betting on one of the dozen groups
3. Betting on the range of numbers from 1 to

18

4. Betting on the range of numbers 19–36
5. Betting on one of the columns
6. Betting on an even or odd number (excluding zero)

Suppose that you have eight cards. Five are green and three are yellow. The five green cards are numbered 1, 2, 3, 4, and 5. The three yellow cards are numbered 1, 2, and 3. The cards are well shuffled. You randomly draw one card.

- $G =$ card drawn is green
- $E =$ card drawn is even-numbered

    1. List the sample space.
    2. $P(G) =$ ____
    3. $P(G|E) =$ ____
    4. $P(G \cap E) =$ ____
    5. $P(G \cup E) =$ ____
    6. Are G and E mutually exclusive? Justify your answer numerically.

---

1. $\{G1, G2, G3, G4, G5, Y1, Y2, Y3\}$
2. 5 8
3. 2 3
4. 2 8
5. 6 8

6. No, because $P(G \cap E)$ does not equal 0.

Roll two fair dice separately. Each die has six faces.

1. List the sample space.
2. Let A be the event that either a three or four is rolled first, followed by an even number. Find $P(A)$.
3. Let B be the event that the sum of the two rolls is at most seven. Find $P(B)$.
4. In words, explain what "$P(A|B)$" represents. Find $P(A|B)$.
5. Are A and B mutually exclusive events? Explain your answer in one to three complete sentences, including numerical justification.
6. Are A and B independent events? Explain your answer in one to three complete sentences, including numerical justification.

A special deck of cards has ten cards. Four are green, three are blue, and three are red. When a card is picked, its color of it is recorded. An experiment consists of first picking a card and then tossing a coin.

1. List the sample space.
2. Let A be the event that a blue card is picked first, followed by landing a head on the coin toss. Find $P(A)$.
3. Let B be the event that a red or green is picked, followed by landing a head on the coin toss. Are the events A and B mutually exclusive? Explain your answer in one to three complete sentences, including numerical justification.
4. Let C be the event that a red or blue is picked, followed by landing a head on the coin toss. Are the events A and C mutually exclusive? Explain your answer in one to three complete sentences, including numerical justification.

NOTE

The coin toss is independent of the card picked first.

1. $\{(G,H)\ (G,T)\ (B,H)\ (B,T)\ (R,H)\ (R,T)\}$
2. $P(A) = P(\text{blue})P(\text{head}) = (\ 3\ 10\ )\ (\ 1\ 2\ )$ $= 3\ 20$
3. Yes, A and B are mutually exclusive

because they cannot happen at the same time; you cannot pick a card that is both blue and also (red or green). $P(A \cap B) = 0$

4. No, A and C are not mutually exclusive because they can occur at the same time. In fact, C includes all of the outcomes of A; if the card chosen is blue it is also (red or blue). $P(A \cap C) = P(A) = 3\ 20$

An experiment consists of first rolling a die and then tossing a coin.

1. List the sample space.
2. Let A be the event that either a three or a four is rolled first, followed by landing a head on the coin toss. Find $P(A)$.
3. Let B be the event that the first and second tosses land on heads. Are the events A and B mutually exclusive? Explain your answer in one to three complete sentences, including numerical justification.

An experiment consists of tossing a nickel, a dime, and a quarter. Of interest is the side the coin lands on.

1. List the sample space.
2. Let A be the event that there are at least

two tails. Find $P(A)$.

3. Let B be the event that the first and second tosses land on heads. Are the events A and B mutually exclusive? Explain your answer in one to three complete sentences, including justification.

---

1. $S = \{(HHH), (HHT), (HTH), (HTT), (THH), (THT), (TTH), (TTT)\}$
2. 4 8
3. Yes, because if A has occurred, it is impossible to obtain two tails. In other words, $P(A \cap B) = 0$.

Consider the following scenario:
Let $P(C) = 0.4$.
Let $P(D) = 0.5$.
Let $P(C|D) = 0.6$.

1. Find $P(C \cap D)$.
2. Are C and D mutually exclusive? Why or why not?
3. Are C and D independent events? Why or why not?
4. Find $P(C \cup D)$.
5. Find $P(D|C)$.

Y and Z are independent events.

1. Rewrite the basic Addition Rule $P(Y \cup Z) = P(Y) + P(Z) - P(Y \cap Z)$ using the information that Y and Z are independent events.
2. Use the rewritten rule to find $P(Z)$ if $P(Y \cup Z) = 0.71$ and $P(Y) = 0.42$.

---

1. If Y and Z are independent, then $P(Y \cap Z) = P(Y)P(Z)$, so $P(Y \cup Z) = P(Y) + P(Z) - P(Y)P(Z)$.
2. 0.5

G and H are mutually exclusive events. $P(G) = 0.5$ $P(H) = 0.3$

1. Explain why the following statement MUST be false: $P(H|G) = 0.4$.
2. Find $P(H \cup G)$.
3. Are G and H independent or dependent events? Explain in a complete sentence.

Approximately 281,000,000 people over age five live in the United States. Of these people, 55,000,000 speak a language other than English at home. Of those who speak another language at home, 62.3% speak Spanish.

Let: $E =$ speaks English at home; $E' =$ speaks another language at home; $S =$ speaks Spanish;

Finish each probability statement by matching the correct answer.

Probability Statements	**Answers**
a. $P(E') =$	i. 0.8043
b. $P(E) =$	ii. 0.623
c. $P(S \cap E') =$	iii. 0.1957
d. $P(S\|E') =$	iv. 0.1219

iii i iv ii

1994, the U.S. government held a lottery to issue 55,000 Green Cards (permits for non-citizens to work legally in the U.S.). Renate Deutsch, from Germany, was one of approximately 6.5 million people who entered this lottery. Let $G =$ won green card.

1. What was Renate's chance of winning a

Green Card? Write your answer as a probability statement.

2. In the summer of 1994, Renate received a letter stating she was one of 110,000 finalists chosen. Once the finalists were chosen, assuming that each finalist had an equal chance to win, what was Renate's chance of winning a Green Card? Write your answer as a conditional probability statement. Let $F =$ was a finalist.
3. Are G and F independent or dependent events? Justify your answer numerically and also explain why.
4. Are G and F mutually exclusive events? Justify your answer numerically and explain why.

Three professors at George Washington University did an experiment to determine if economists are more selfish than other people. They dropped 64 stamped, addressed envelopes with $10 cash in different classrooms on the George Washington campus. 44% were returned overall. From the economics classes 56% of the envelopes were returned. From the business, psychology, and history classes 31% were returned.

Let: $R =$ money returned; $E =$ economics classes; $O =$ other classes

1. Write a probability statement for the overall percent of money returned.
2. Write a probability statement for the percent of money returned out of the economics classes.
3. Write a probability statement for the percent of money returned out of the other classes.
4. Is money being returned independent of the class? Justify your answer numerically and explain it.
5. Based upon this study, do you think that economists are more selfish than other people? Explain why or why not. Include numbers to justify your answer.

---

1. $P(R) = 0.44$
2. $P(R|E) = 0.56$
3. $P(R|O) = 0.31$
4. No, whether the money is returned is not independent of which class the money was placed in. There are several ways to justify this mathematically, but one is that the money placed in economics classes is not returned at the same overall rate; $P(R|E) \neq P(R)$.
5. No, this study definitely does not support that notion; *in fact*, it suggests the opposite. The money placed in the economics classrooms was returned at a

higher rate than the money place in all classes collectively; $P(R|E) > P(R)$.

The following table of data obtained from www.baseball-almanac.com shows hit information for four players. Suppose that one hit from the table is randomly selected.

Name	Single	Double	Triple	Home run	Total hits
Babe Ruth	1,517	506	136	714	2,873
Jackie Robinson	1,054	273	54	137	1,518
Ty Cobb	3,603	174	295	114	4,189
Hank Aaron	2,294	624	98	755	3,771
Total	8,471	1,577	583	1,720	12,351

Are "the hit being made by Hank Aaron" and "the hit being a double" independent events?

1. Yes, because $P(\text{hit by Hank Aaron}|\text{hit is a double}) = P(\text{hit by Hank Aaron})$

2. No, because P(hit by Hank Aaron|hit is a double) $\neq$ P(hit is a double)
3. No, because P(hit is by Hank Aaron|hit is a double) $\neq$ P(hit by Hank Aaron)
4. Yes, because P(hit is by Hank Aaron|hit is a double) $=$ P(hit is a double)

United Blood Services is a blood bank that serves more than 500 hospitals in 18 states. According to their website, a person with type O blood and a negative Rh factor (Rh-) can donate blood to any person with any bloodtype. Their data show that 43% of people have type O blood and 15% of people have Rh- factor; 52% of people have type O or Rh- factor.

1. Find the probability that a person has both type O blood and the Rh- factor.
2. Find the probability that a person does NOT have both type O blood and the Rh- factor.

---

1. P(type O $\cup$ Rh-) $=$ P(type O) $+$ P(Rh-) - P(type O $\cap$ Rh-)

   0.52 $=$ 0.43 $+$ 0.15 - P(type O $\cap$ Rh-); solve to find P(type O $\cap$ Rh-) $=$ 0.06

   6% of people have type O, Rh- blood

2. *P*(NOT(type O ∩ Rh-)) = 1 - *P*(type O ∩ Rh-) = 1 - 0.06 = 0.94

   94% of people do not have type O, Rh- blood

At a college, 72% of courses have final exams and 46% of courses require research papers. Suppose that 32% of courses have a research paper and a final exam. Let *F* be the event that a course has a final exam. Let *R* be the event that a course requires a research paper.

1. Find the probability that a course has a final exam or a research project.
2. Find the probability that a course has NEITHER of these two requirements.

In a box of assorted cookies, 36% contain chocolate and 12% contain nuts. Of those, 8% contain both chocolate and nuts. Sean is allergic to both chocolate and nuts.

1. Find the probability that a cookie contains chocolate or nuts (he can't eat it).
2. Find the probability that a cookie does not contain chocolate or nuts (he can eat it).

---

1. Let C = be the event that the cookie contains chocolate. Let N = the event that the cookie contains nuts.
2. $P(C \cup N) = P(C) + P(N) - P(C \cap N) = 0.36 + 0.12 - 0.08 = 0.40$
3. P(NEITHER chocolate NOR nuts) = $1 - P(C \cup N) = 1 - 0.40 = 0.60$

A college finds that 10% of students have taken a distance learning class and that 40% of students are part time students. Of the part time students, 20% have taken a distance learning class. Let D = event that a student takes a distance learning class and E = event that a student is a part time student

1. Find $P(D \cap E)$.
2. Find $P(E|D)$.
3. Find $P(D \cup E)$.
4. Using an appropriate test, show whether D and E are independent.
5. Using an appropriate test, show whether D and E are mutually exclusive.

Glossary

Independent Events

The occurrence of one event has no effect on the probability of the occurrence of another event. Events A and B are independent if one of the following is true:

1. $P(A|B) = P(A)$
2. $P(B|A) = P(B)$
3. $P(A \cap B) = P(A)P(B)$

Mutually Exclusive

Two events are mutually exclusive if the probability that they both happen at the same time is zero. If events A and B are mutually exclusive, then $P(A \cap B) = 0$.

Contingency Tables and Tree Diagrams

Contingency Tables

A **contingency table** provides a way of portraying data that can facilitate calculating probabilities. The table helps in determining conditional probabilities quite easily. The table displays sample values in relation to two different variables that may be dependent or contingent on one another. Later on, we will use contingency tables again, but in another manner.

Suppose a study of speeding violations and drivers who use cell phones produced the following fictional data:

	Speeding violation in the last year	**No speeding violation in the last year**	**Total**

Cell phone user	25	280	305
Not a cell phone user	45	405	450
Total	70	685	755

The total number of people in the sample is 755. The row totals are 305 and 450. The column totals are 70 and 685. Notice that 305 + 450 = 755 and 70 + 685 = 755.
Calculate the following probabilities using the table.

a. Find *P*(Person is a car phone user).

a. number of car phone users
total number in study = 305 755

b. Find *P*(person had no violation in the last year).

b. number that had no violation
total number in study = 685 755

c. Find *P*(Person had no violation in the last year AND was a car phone user).

c. 280 755

d. Find *P*(Person is a car phone user OR person had no violation in the last year).

d. (305 755 + 685 755) − 280 755 = 710 755

e. Find *P*(Person is a car phone user GIVEN person had a violation in the last year).

e. 25 70 (The sample space is reduced to the number of persons who had a violation.)

f. Find *P*(Person had no violation last year GIVEN person was not a car phone user)

f. 405 450 (The sample space is reduced to the number of persons who were not car phone users.)

Try it

[link] shows the number of athletes who stretch before exercising and how many had injuries within the past year.

	Injury in last year	**No injury in last year**	**Total**
Stretches	55	295	350
Does not stretch	231	219	450
Total	286	514	800

1. What is *P*(athlete stretches before exercising)?
2. What is *P*(athlete stretches before exercising|no injury in the last year)?

1. *P*(athlete stretches before exercising) =

350 800 = 0.4375

2. P(athlete stretches before exercising|no injury in the last year) = 295 514 = 0.5739

[link] shows a random sample of 100 hikers and the areas of hiking they prefer.

Sex	The Coastline	Near Lakes and Streams	On Mountain Peaks	Total
Female	18	16	___	45
Male	___	___	14	55
Total	___	41	___	___

Hiking Area Preference

a. Complete the table.

a.

Sex	The Coastline	Near Lakes and Streams	On Mountain Peaks	Total
Female	18	16	11	45
Male	16	25	14	55
Total	34	41	25	100

Hiking Area Preference

b. Are the events "being female" and "preferring the coastline" independent events?

Let F = being female and let C = preferring the coastline.

1. Find $P(F \text{ AND } C)$.
2. Find $P(F)P(C)$

Are these two numbers the same? If they are, then F and C are independent. If they are not, then F and C are not independent.

b.

1. $P(F \text{ AND } C) = \frac{18}{100} = 0.18$
2. $P(F)P(C) = \left(\frac{45}{100}\right)\left(\frac{34}{100}\right) = (0.45)(0.34) = 0.153$

$P(F \text{ AND } C) \neq P(F)P(C)$, so the events F and C

are not independent.

c. Find the probability that a person is male given that the person prefers hiking near lakes and streams. Let $M =$ being male, and let $L =$ prefers hiking near lakes and streams.

1. What word tells you this is a conditional?
2. Fill in the blanks and calculate the probability: $P(\_\_|\_\_) = \_\_$.
3. Is the sample space for this problem all 100 hikers? If not, what is it?

---

c.

1. The word 'given' tells you that this is a conditional.
2. $P(M|L) = 25\ 41$
3. No, the sample space for this problem is the 41 hikers who prefer lakes and streams.

d. Find the probability that a person is female or prefers hiking on mountain peaks. Let $F =$ being female, and let $P =$ prefers mountain peaks.

1. Find $P(F)$.
2. Find $P(P)$.
3. Find $P(F \text{ AND } P)$.
4. Find $P(F \text{ OR } P)$.

d.

1. $P(F) = 45\ 100$
2. $P(P) = 25\ 100$
3. $P(F \text{ AND } P) = 11\ 100$
4. $P(F \text{ OR } P) = 45\ 100 + 25\ 100 \text{ - } 11\ 100$
   $= 59\ 100$

Try It

[link] shows a random sample of 200 cyclists and the routes they prefer. Let M = males and H = hilly path.

Gender	Lake	Hilly	Wooded	Total

	Path	Path	Path	
Female	45	38	27	110
Male	26	52	12	90
Total	71	90	39	200

1. Out of the males, what is the probability that the cyclist prefers a hilly path?
2. Are the events “being male” and “preferring the hilly path” independent events?

1. $P(H|M) = 52\ 90 = 0.5778$
2. For M and H to be independent, show $P(H|M) = P(H)$

   $P(H|M) = 0.5778$, $P(H) = 90\ 200 = 0.45$

   $P(H|M)$ does not equal $P(H)$ so M and H are NOT independent.

Muddy Mouse lives in a cage with three doors. If Muddy goes out the first door, the probability that he gets caught by Alissa the cat is 1 5 and the probability he is not caught is 4 5 . If he goes out the second door, the probability he gets caught by Alissa is 1 4 and the probability he is not caught is 3 4 . The probability that Alissa catches Muddy coming out of the third door is 1 2 and the probability she does not catch Muddy is 1 2 . It is equally likely that Muddy will choose any of the three doors so the probability of choosing each door is 1 3 .

Caught or Not	**Door One**	**Door Two**	**Door Three**	**Total**
Caught	1 15	1 12	1 6	___-
Not Caught	4 15	3 12	1 6	___
Total	___	___	___	1

Door Choice

- The first entry 1 15 = (1 5) (1 3) is *P*(Door One AND Caught)
- The entry 4 15 = (4 5)(1 3) is *P*(Door One AND Not Caught)

Verify the remaining entries.

a. Complete the probability contingency table. Calculate the entries for the totals. Verify that the lower-right corner entry is 1.

a.

Caught or Not	**Door One**	**Door Two**	**Door Three**	**Total**
Caught	1 15	1 12	1 6	**1960**
Not Caught	4 15	3 12	1 6	**4160**
Total	**515**	**412**	**26**	1

Door Choice

b. What is the probability that Alissa does not catch Muddy?

b. 4160

c. What is the probability that Muddy chooses Door One OR Door Two given that Muddy is caught by Alissa?

c. 919

[link] contains the number of crimes per 100,000 inhabitants from 2008 to 2011 in the U.S.

Year	Robbery	Burglary	Rape	Vehicle	Total
2008	145.7	732.1	29.7	314.7	
2009	133.1	717.7	29.1	259.2	
2010	119.3	701	27.7	239.1	
2011	113.7	702.2	26.8	229.6	
Total					

United States Crime Index Rates Per 100,000 Inhabitants 2008–2011

TOTAL each column and each row. Total data $= 4,520.7$

1. Find P(2009 AND Robbery).
2. Find P(2010 AND Burglary).
3. Find P(2010 OR Burglary).
4. Find $P(2011|\text{Rape})$.
5. Find $P(\text{Vehicle}|2008)$.

a. 0.0294, b. 0.1551, c. 0.7165, d. 0.2365, e. 0.2575

Try It

[link] relates the weights and heights of a group of individuals participating in an observational study.

Weight/Height	Tall	Medium	Short	Totals
Obese	18	28	14	
Normal	20	51	28	
Underweight	12	25	9	
Totals				

1. Find the total for each row and column
2. Find the probability that a randomly chosen individual from this group is Tall.
3. Find the probability that a randomly chosen individual from this group is Obese and Tall.
4. Find the probability that a randomly chosen individual from this group is Tall given that the idividual is Obese.
5. Find the probability that a randomly chosen individual from this group is Obese given that the individual is Tall.
6. Find the probability a randomly chosen

individual from this group is Tall and Underweight.

7. Are the events Obese and Tall independent?

Weight/ Height	Tall	Medium	Short	Totals
Obese	18	28	14	60
Normal	20	51	28	99
Underweight	12	25	9	46
Totals	50	104	51	205

1. Row Totals: 60, 99, 46. Column totals: 50, 104, 51.
2. P(Tall) = 50 205 = 0.244
3. P(Obese AND Tall) = 18 205 = 0.088
4. P(Tall|Obese) = 18 60 = 0.3
5. P(Obese|Tall) = 18 50 = 0.36
6. P(Tall AND Underweight = 12 205 = 0.0585
7. No. P(Tall) does not equal P(Tall|Obese).

Tree Diagrams

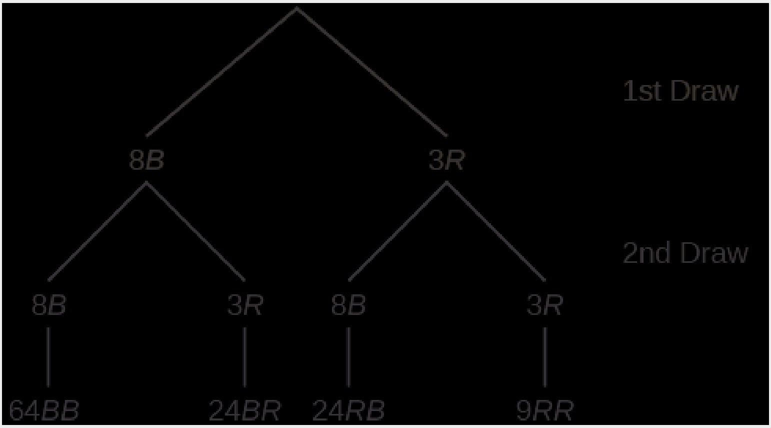
Sometimes, when the probability problems are complex, it can be helpful to graph the situation. Tree diagrams can be used to visualize and solve conditional probabilities.

Tree Diagrams

A **tree diagram** is a special type of graph used to determine the outcomes of an experiment. It consists of "branches" that are labeled with either frequencies or probabilities. Tree diagrams can make some probability problems easier to visualize and solve. The following example illustrates how to use a tree diagram.

In an urn, there are 11 balls. Three balls are red (R) and eight balls are blue (B). Draw two balls, one at a time, **with replacement**. "With replacement" means that you put the first ball back in the urn before you select the second ball. The tree diagram using frequencies that show all the possible outcomes follows.

$\text{Total} = 64 + 24 + 24 + 9 = 121$



The first set of branches represents the first draw. The second set of branches represents the second draw. Each of the outcomes is distinct. In fact, we can list each red ball as *R*1, *R*2, and *R*3 and each blue ball as *B*1, *B*2, *B*3, *B*4, *B*5, *B*6, *B*7, and *B*8. Then the nine *RR* outcomes can be written as:
*R*1*R*1 *R*1*R*2 *R*1*R*3 *R*2*R*1 *R*2*R*2 *R*2*R*3 *R*3*R*1 *R*3*R*2 *R*3*R*3

The other outcomes are similar.

There are a total of 11 balls in the urn. Draw two balls, one at a time, with replacement. There are 11(11) = 121 outcomes, the size of the **sample space**.

a. List the 24 *BR* outcomes: *B*1*R*1, *B*1*R*2, *B*1*R*3, ...

a. *B*1*R*1 *B*1*R*2 *B*1*R*3 *B*2*R*1 *B*2*R*2 *B*2*R*3 *B*3*R*1 *B*3*R*2 *B*3*R*3 *B*4*R*1 *B*4*R*2 *B*4*R*3 *B*5*R*1 *B*5*R*2 *B*5*R*3 *B*6*R*1 *B*6*R*2 *B*6*R*3 *B*7*R*1 *B*7*R*2 *B*7*R*3

*B*8*R*1 *B*8*R*2 *B*8*R*3

b. Using the tree diagram, calculate *P*(*RR*).

b. *P*(*RR*) = (3 11)(3 11) = 9 121

c. Using the tree diagram, calculate *P*(*RB* OR *BR*).

c. *P*(*RB* OR *BR*) = (3 11)(8 11) + (8 11)(3 11) = 48 121

d. Using the tree diagram, calculate *P*(*R* on 1st draw AND *B* on 2nd draw).

d. *P*(*R* on 1st draw AND *B* on 2nd draw) = *P*(*RB*) = (3 11)(8 11) = 24 121

e. Using the tree diagram, calculate *P*(*R* on 2nd draw GIVEN *B* on 1st draw).

e. $P(R$ on 2nd draw GIVEN B on 1st draw$)$ = $P(R$ on 2nd$|B$ on 1st$) = 24\ 88 = 3\ 11$

This problem is a conditional one. The sample space has been reduced to those outcomes that already have a blue on the first draw. There are $24 + 64 = 88$ possible outcomes (24 *BR* and 64 *BB*). Twenty-four of the 88 possible outcomes are *BR*. $24\ 88 = 3\ 11$.

f. Using the tree diagram, calculate $P(BB)$.

f. $P(BB) = \ 64\ 121$

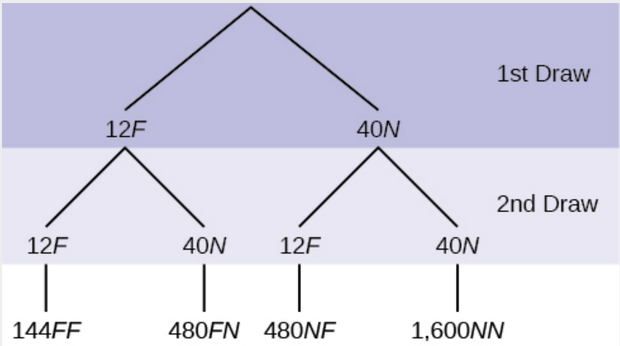
g. Using the tree diagram, calculate $P(B$ on the 2nd draw given R on the first draw$)$.

g. $P(B$ on 2nd draw$|R$ on 1st draw$) = \ 8\ 11$

There are $9 + 24$ outcomes that have R on the first draw (9 *RR* and 24 *RB*). The sample space is then $9 + 24 = 33$. 24 of the 33 outcomes have B on the second draw. The probability is then 24 33 .

Try It

In a standard deck, there are 52 cards. 12 cards are face cards (event *F*) and 40 cards are not face cards (event *N*). Draw two cards, one at a time, with replacement. All possible outcomes are shown in the tree diagram as frequencies. Using the tree diagram, calculate *P*(*FF*).



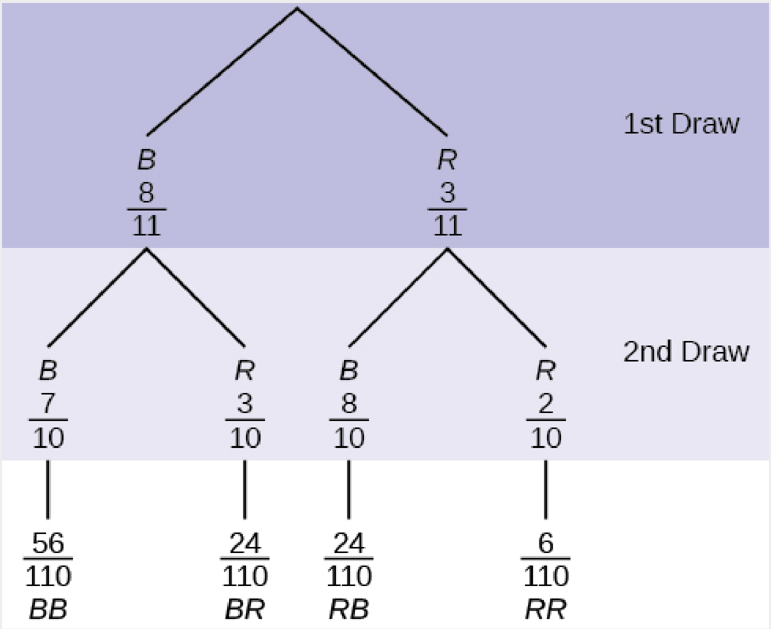
Total number of outcomes is 144 + 480 + 480 + 1600 = 2,704.

P(*FF*) = 144 144 + 480 + 480 + 1,600 = 144 2,704 = 9 169

An urn has three red marbles and eight blue marbles in it. Draw two marbles, one at a time, this time without replacement, from the urn. "**Without**

replacement" means that you do not put the first ball back before you select the second marble. Following is a tree diagram for this situation. The branches are labeled with probabilities instead of frequencies. The numbers at the ends of the branches are calculated by multiplying the numbers on the two corresponding branches, for example, (3 11)(2 10)= 6 110 .

Total = 56+24+24+6 110 = 110 110 =1



NOTE

If you draw a red on the first draw from the three red possibilities, there are two red marbles left to draw on the second draw. You do not put back or replace the first marble after you have drawn it.

You draw **without replacement**, so that on the second draw there are ten marbles left in the urn.

Calculate the following probabilities using the tree diagram.

a. $P(RR)$ = ______

a. $P(RR)$ = (3 11)(2 10)= 6 110

b. Fill in the blanks:

$P(RB$ OR $BR)$ = (3 11)(8 10) + (__)(__) = 48 110

b. $P(RB$ OR $BR)$ = (3 11)(8 10) + (8 11)(3 10) = 48 110

c. $P(R$ on 2nd$|B$ on 1st$)$ =

c. $P(R$ on 2nd$|B$ on 1st$)$ = 3 10

d. Fill in the blanks.

$P(R \text{ on 1st AND } B \text{ on 2nd}) = P(RB) = (\_\_)(\_\_) = 24\ 100$

d. $P(R \text{ on 1st AND } B \text{ on 2nd}) = P(RB) = (\ 3\ 11\ )(\ 8\ 10\ ) = 24\ 100$

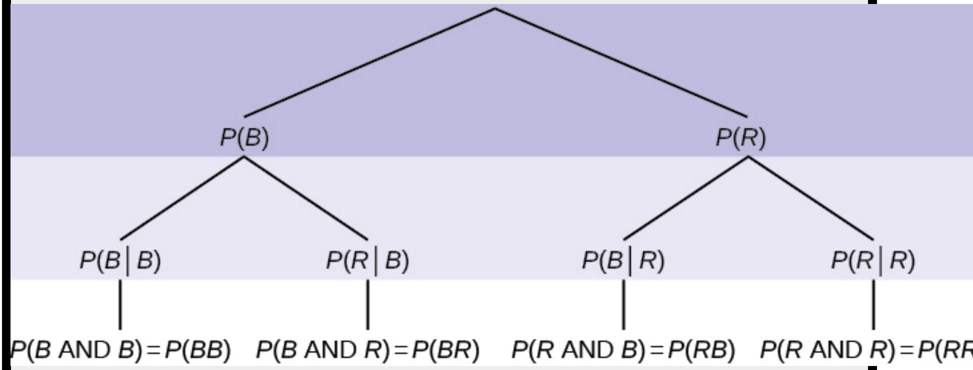
e. Find $P(BB)$.

e. $P(BB) = (\ 8\ 11\ )(\ 7\ 10\ )$

f. Find $P(B \text{ on 2nd} | R \text{ on 1st})$.

f. Using the tree diagram, $P(B \text{ on 2nd} | R \text{ on 1st}) = P(R|B) = 8\ 10$.

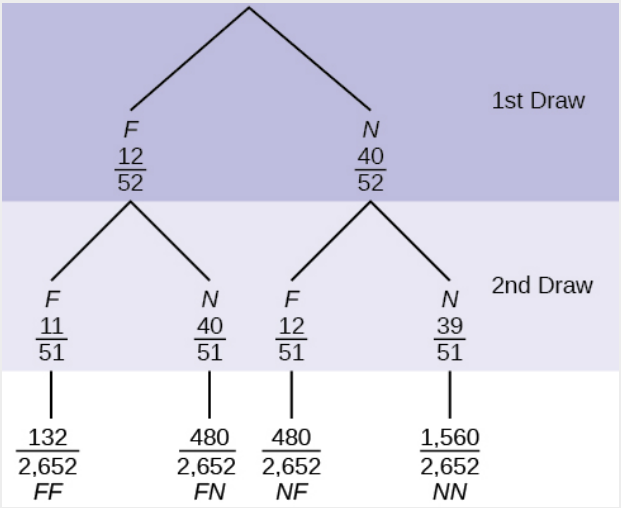
If we are using probabilities, we can label the tree in the following general way.



- $P(R|R)$ here means $P(R \text{ on 2nd}|R \text{ on 1st})$
- $P(B|R)$ here means $P(B \text{ on 2nd}|R \text{ on 1st})$
- $P(R|B)$ here means $P(R \text{ on 2nd}|B \text{ on 1st})$
- $P(B|B)$ here means $P(B \text{ on 2nd}|B \text{ on 1st})$

Try It

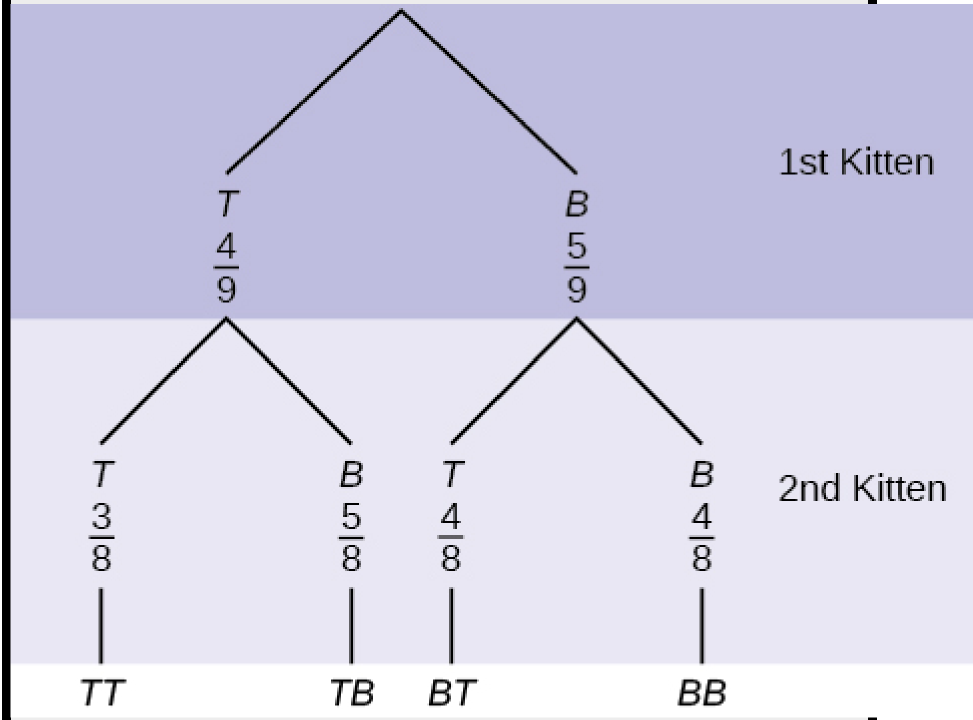
In a standard deck, there are 52 cards. Twelve cards are face cards (*F*) and 40 cards are not face cards (*N*). Draw two cards, one at a time, without replacement. The tree diagram is labeled with all possible probabilities.



1. Find *P*(*FN* OR *NF*).
2. Find *P*(*N*|*F*).
3. Find *P*(at most one face card).
   Hint: "At most one face card" means zero or one face card.
4. Find *P*(at least on face card).
   Hint: "At least one face card" means one or two face cards.

1. *P*(*FN* OR *NF*) = 480 2,652 + 480 2,652 = 960 2,652 = 80 221
2. *P*(*N*|*F*) = 40 51
3. *P*(at most one face card) = (480 + 480 + 1,560) 2,652 = 2,520 2,652
4. *P*(at least one face card) = (132 + 480 + 480) 2,652 = 1,092 2,652

A litter of kittens available for adoption at the Humane Society has four tabby kittens and five black kittens. A family comes in and randomly selects two kittens (without replacement) for adoption.



1. What is the probability that both kittens are tabby?

   a. (1 2)(1 2) b. (4 9)(4 9) c. (4 9)(3 8) d. (4 9)(5 9)
2. What is the probability that one kitten of

each coloring is selected?

a. (4 9)(5 9) b. (4 9)(5 8) c. (4 9)(5 9)+(5 9)(4 9) d. (4 9)(5 8)+(5 9)(4 8)

3. What is the probability that a tabby is chosen as the second kitten when a black kitten was chosen as the first?
4. What is the probability of choosing two kittens of the same color?

a. c, b. d, c. 4 8 , d. 32 72

Try It

Suppose there are four red balls and three yellow balls in a box. Three balls are drawn from the box without replacement. What is the probability that one ball of each coloring is selected?

(4 7)(3 6) + (3 7)(4 6)

References

"Blood Types." American Red Cross, 2013. Available online at http://www.redcrossblood.org/learn-about-blood/blood-types (accessed May 3, 2013).

Data from the National Center for Health Statistics, part of the United States Department of Health and Human Services.

Data from United States Senate. Available online at www.senate.gov (accessed May 2, 2013).

"Human Blood Types." Unite Blood Services, 2011. Available online at http://www.unitedbloodservices.org/learnMore.aspx (accessed May 2, 2013).

Haiman, Christopher A., Daniel O. Stram, Lynn R. Wilkens, Malcom C. Pike, Laurence N. Kolonel, Brien E. Henderson, and Loīc Le Marchand. "Ethnic and Racial Differences in the Smoking-Related Risk of Lung Cancer." The New England Journal of Medicine, 2013. Available online at http://www.nejm.org/doi/full/10.1056/NEJMoa033250 (accessed May 2, 2013).

Samuel, T. M. "Strange Facts about RH Negative Blood." eHow Health, 2013. Available online at http://www.ehow.com/facts_5552003_strange-rh-negative-blood.html (accessed May 2, 2013).

“United States: Uniform Crime Report – State Statistics from 1960–2011.” The Disaster Center. Available online at http://www.disastercenter.com/crime/ (accessed May 2, 2013).

Data from Clara County Public H.D.

Data from the American Cancer Society.

Data from The Data and Story Library, 1996. Available online at http://lib.stat.cmu.edu/DASL/ (accessed May 2, 2013).

Data from the Federal Highway Administration, part of the United States Department of Transportation.

Data from the United States Census Bureau, part of the United States Department of Commerce.

Data from USA Today.

“Environment.” The World Bank, 2013. Available online at http://data.worldbank.org/topic/environment (accessed May 2, 2013).

“Search for Datasets.” Roper Center: Public Opinion Archives, University of Connecticut., 2013. Available online at http://www.ropercenter.uconn.edu/data_access/data/search_for_datasets.html (accessed May 2, 2013).

Chapter Review

There are several tools you can use to help organize and sort data when calculating probabilities. Contingency tables help display data and are particularly useful when calculating probabilites that have multiple dependent variables.

A tree diagram use branches to show the different outcomes of experiments and makes complex probability questions easy to visualize.

Glossary

Tree Diagram

the useful visual representation of a sample space and events in the form of a “tree” with branches marked by possible outcomes together with associated probabilities (frequencies, relative frequencies)

Contingency Table

the method of displaying a frequency distribution as a table with rows and columns to show how two variables may be dependent (contingent) upon each other; the table provides an easy way to calculate conditional probabilities.

Introduction

Chapter Objectives

By the end of this chapter, the student should be able to:

- Recognize the binomial probability distribution and apply it appropriately.
- Be able to evaluate evidence using the binomial distribution.

Suppose you flip a coin ten times and each time it comes up heads. This might make you start to wonder if there is something wrong with the coin. Perhaps it is a trick coin and is heads on both sides? Perhaps it is imbalanced and it is more likely to come up heads over tails. You may also wonder what is the probability of getting 10 heads in a row, if the coin was fair.

Coin flipping is interesting because it is a random event. We cannot predict whether the next flip will be heads or tails (assuming it isn’t a trick coin). That means the outcome would be a random variable. A **random variable** is any variable where the outcome is determined by a random event. The outcome is also discrete because we count it. Above you flipped a coin ten times and counted the number of heads. A **discrete random variable** is a variable whose outcome is determined by a random event and where we count the outcomes. Other examples of discrete random variables include how many times you roll an even number with a die out of ten rolls; how many customers enter a store during a five-minute interval; how many times you draw a high card out of a deck of cards out of eight draws (without replacement).

In each of these situations (coin toss, rolling die, number of customers, drawing cards), you could look at each situation and, each time, come up with a new formula to find the probability of these events happening. But this would take a lot of work and be inefficient. Instead, you would want to see if the situation can be **modelled** by a distribution. A **probability distribution** provides the theoretical probabilities of all of the possible events in a situation. For example, the following is a probability model of how many heads you can get when you flip a fair coin three times:

# of heads		Probability

0	0.125
1	0.375
2	0.375
3	0.125

Notice that the probabilities range from 0 to 1 and that the sum of the probabilities is 1.

The above table could be determined by working out all of the possible outcomes (TTT, TTH, THT, HTT, etc.), then counting how many heads were in each outcome. But again, that is time consuming. Instead, you want to see if there is a probability distribution that models your situation that you can use. For example, coin tossing can be modelled by the T distribution.

The binomial distribution is an example of a model for discrete random variables. There are many other models for discrete random variables including Poisson, geometric, hypergeometric, and discrete uniform to name a few. Each distribution comes with a set of criteria and if a situation fits that criteria, then the distribution can model it. That is, the distribution can produce theoretical probabilities for that situation.

Theoretical vs. experimental probabilities

In simplest terms, a theoretical probability is determined by using a formula while an experimental probability is found by actually doing the event. For example, if you flip a coin 3 times and get 2 heads, then the experimental probability is 2/3 = 0.6667. The theoretical probability is 0.375. The theoretical probability is also called the **long-run probability**, because the longer you do the experimental probability the closer the experimental results will get to the theoretical probability. This is an example of the **law of large numbers**.

In this chapter, we are going to learn about the binomial distribution, which is a model for discrete random variables. In the next chapter, we will learn about the normal distribution, which is a model for continuous random variables.

In particular, we want to use the binomial distribution to evaluate evidence. For example, going back to the example flipping a coin ten times and getting ten heads, we want to use the evidence (getting ten heads) to determine whether we think there is something wrong with the coin.

The Binomial Distribution
An introduction to the binomial distribution an informal inferential statistics.

Binomial distribution

Flipping a coin a certain number of times, let's say ten times, is a classic example of a binomial distribution. What are the characteristics of flipping a coin that makes it binomial?

Before we answer that question, let's get a bit terminology out of the way. In probability theory, an **experiment** is the actual process that you investigating. In the flipping coin example, the experiment is flipping a coin ten times. A **trial** is a specific instance of an experiment. Flipping a coin only once is considered a trial.

Going back to the coin flipping example, let's assume that we are dealing with a fair coin (i.e. probability of getting a head or a tail is 50%). We've already discussed that when we count the number of heads that this is an example of a discrete random variable. Notice that there are only two possible outcomes (heads or tails). This is a key criterion for a binomial distribution (*binomial* derives from Latin for two terms). Also, notice that the events are independent of each other. That is, if you get a head on one flip, that has no impact on the probability of getting a head on the next flip. This also means that the probability of getting a head remains constant. This is another key criterion for a binomial distribution. The other thing to notice is that the number of times we flip the coin is fixed. We don't flip it until we get bored or run out of time. Instead we flip it ten times. This means that the number of trials is fixed. This is the last key criterion of a binomial distribution.

There are five characteristics of a binomial experiment.

1. The variable being studied is random.
2. The outcomes of the variable are being counted.
3. There are a fixed number of trials. The letter *n* denotes the number of trials.
4. There are only two possible outcomes, called "success" and "failure," for each trial π denotes the probability of a success on one trial, and 1- π denotes the probability of a failure on one trial.
5. The *n* trials are independent and are repeated using identical conditions. Because the *n* trials are independent, the outcome of one trial does not help in predicting the outcome of another trial. Another way of saying this is that for each individual trial, the probability, π , of a success and probability, 1- π , of a failure remain

constant.

Other examples of binomial distributions:

- Counting the number of 2's that are rolled when you roll a die six times (trial = rolling a dice, success = rolling a 2, n = 6, π = 1/6 = 0.1667)
- Counting the number of times a jack is pulled out of deck of cards (with replacement) when you pull a card fifteen times (trial = pulling a card, success = pulling a jack, n = 15, π = 4/52 = 0.0769)
- Counting the number of times that you win a prize in Tim Hortons Roll up the Rim to Win contest out of four cups (trial = checking cup for win, success = winning a prize, n = 4, π = 1/6 = 0.1667 – assuming no special rules (e.g. anniversary rules that changed the odds of winning)

Examples of situations that are not binomial include:

- Counting the number of times a jack is pulled out of deck of cards (without replacement) when you pull a card fifteen times. The fifth criterion is not met because the events are now dependent.
- Counting the number times each number (1 to 6) is rolled when you roll a die fifty times. The

fourth criterion is not met because there are six possible outcomes instead of two.

- Counting the number of times that you win a prize in Tim Horton's Roll up the Rim to Win contest out of how many cups you buy during the contest. Unless you know exactly how many you'll buy during the contest, this would not meet the third criterion of having a fixed number of trials.

The Roll up the Rim example might not be binomial as it may fail the fifth criterion. At the beginning of the contest, the odds of winning are determined by counting how many prizes there are out of the total number of cups printed. As the contest goes on, the probability of winning may change depending on how many people have already won. At the beginning of the contest, this is also true but there are so many cups that it doesn't really matter (think back to the sampling with replacement vs. without replacement in Chapter 1). Thus, this contest is only binomial at the beginning of the contest.

Notation

Suppose we are working on a probability question and there are multiple probabilities that need to be found. Then it gets time consuming to write out, for example, “the probability that three rolls of a die will result in at least one 2” or some variation over and over again. Instead we will use notation to reduce the work. We can write the previous statement more quickly as $P(X \geq 1)$. The $P(\ )$ means the “probability of”. X is the random variable being studied (in this case the number of times 2 has been rolled out of 3 rolls). “$X \geq 1$” means we are looking at the number of times a 2 is rolled at least once.

It is important to define X. Otherwise, $P(2 < X \leq 5)$ could refer to any random variable and the person reading the notation won’t know what it means.

Mean and standard deviation of the binomial distribution

Just like a set of data, a binomial distribution has a mean and a standard deviation. For the binomial distribution, these are given by the formulas:

$\mu = n\pi$

$\sigma = n\pi\,(1 - \pi\,)$

Going back to the Tim Hortons example, we had $n = 4$ and $\pi = 0.1667$. Thus $\mu = 4 \times 0.1667 = 0.6667$ and $\sigma = 4 \times 0.1667 \times (1 - 0.1667) = 0.745$.

This means that if we buy four random cups of Tim Hortons coffee during the Roll Up the Rim content, we will typically win 0.67 times, give or take 0.75. Thus, when buying four cups of coffee, we will typically win between -0.08 and 1.42 times. Since we can't win negative times, we will round the lower bound to 0. Therefore, when buying four cups of coffee, we will typically win between 0 and 1.42 times.

> A market research study shows that 30% of all passengers on Canadian Airlines are business travelers. A random sample of 20 passengers is taken.
>
> 1. Explain why the above situation satisfies the criteria of a binomial distribution. If there are any issues with why this situation may not meet all of the criteria, discuss them. Define n, X and π .
> 2. For the random sample, determine the probability that:
>
> 1. Exactly seven of the passengers are business travelers.
> 2. From ten to fourteen (inclusive) are business travelers.
> 3. At least eleven of these passengers are business travelers.
> 4. Five of these passengers are NOT travelling on business.

1. What is the typical range of business passengers in a random sample of 20?

---

1. The situation is a binomial distribution because:

- It represents a random variable as the sample is randomly selected.
- It is a discrete variable as we are counting the number of business travellers. X is the number of business travellers in the sample.
- There is a fixed number of trials ($n = 20$)
- There are only two options: the passenger is a business traveller (success) or they are not a business traveller (failure).
- In a random sample whether one passenger is a business traveller does not affect the probability of another passenger being a business traveller. Therefore, the probability of success remains constant: $\pi = 30\% = 0.3$

1. Use a computer program to come up with the following output.

X	P(X)	P(X ≤ x)
0	0.00080	0.00080
1	0.00684	0.00764
2	0.02785	0.03548
3	0.07160	0.10709
4	0.13042	0.23751
5	0.17886	0.41637
6	0.19164	0.60801
7	0.16426	0.77227
8	0.11440	0.88667
9	0.06537	0.95204
10	0.03082	0.98286
11	0.01201	0.99486
12	0.00386	0.99872
13	0.00102	0.99974
14	0.00022	0.99996
15	0.00004	0.99999
16	0.00001	1.00000
17	0.00000	1.00000
18	0.00000	1.00000
19	0.00000	1.00000
20	0.00000	1.00000

1. P(X=7) = 0.16426
2. P(10 ≤ X ≤ 14) = 0.04792 (highlight the values in the column *P*(*X*) for *X* from 10 to 14, then look at the Sum in the lower right)
3. P(X ≥ 11) = 0.01714 (highlight the values in the column *P*(*X*) for *X* from 11 and higher, then look at the Sum in the

lower right)

4. This changes π to 0.7, then re-run the computer program. Look at when X is 5. $P(X=5) = 0.00004$

1. The mean is the same as the expected value, which is 6.0 and the standard deviation is 2.049. This gives us a typical range of 3.951 and 8.049 for the typical number of business passengers in a random sample of 20 passengers.

Evaluating evidence using the binomial distribution

A company looked at its hiring practices. In particular, they found that their hiring practices appears to favour men over women. Based on past data, they have found that regardless of the number of applications by women, seventy-five percent of hires are men. Due to this issue, they decide to implement program. In this program, the name and any identifying features that may indicate the gender of an applicant are removed. For example, if the application says, “She executed a marketing campaign that increased revenue by 30%”, this would be changed to “They executed a marketing campaign that increased revenue by 30%.” The names on the applications were changed an alpha-numeric identification (like AB-101). The company *claims* that the program has worked, but they want to check the claim.

How will the company determine if the program has worked? One way to do this would be using statistics.

Now suppose that after a recent round of hiring, the proportion of men hired was 70%. Would this be enough evidence that the program is working? 70% is definitely lower than 75%, but we know that there is variability in sampling. This means that, prior to the program being implemented, around 75% of hires are men, there may be some rounds of hiring where 70% of hires were men and some that were 80%. It won't be 75% each time. Instead we expect it to be close to 75%. Therefore, if the program has caused the hiring practices to change, would a recent round of hiring that results in the proportion of men hires being 70% be enough evidence of change? What about 60%? What is the line between normal variability from 75% and abnormal variability? Statistics helps us figure that out and that is how we evaluate evidence using statistics.

Let's say in a recent round of hires there were 30 new hires and 20 of these hires were men.

Skepticism

Any time we are trying to evaluate evidence, we always start from a position of skepticism. That is, we don't want to assume what we are trying to show (i.e the claim). If we do that, we may bias the investigation. To illustrate, if you assume that your significant other is cheating on you, then this will colour all of the evidence you find (why did they show up five minutes late from work? They must be cheating!). A well-known real-world example of this position is the assumption in court that a defendant is innocent until proven guilty. That is, criminal court cases start with the assumption of innocence.

In general, the position of skepticism is that nothing has changed, the program didn't work, the experiment didn't work, the effect being studied isn't happening, etc.

In our example, we will assume that the program that the company implemented did not work. That is, we are assuming that the proportion of hires that are men is still 75%. Another way of writing this is $\pi = 75\%$ (i.e. the population proportion).

Evidence

In a court case, evidence would be witness testimony, forensics evidence, expert testimony, etc.

In statistics, evidence is sample data. The evidence has been collected to evaluate the claim. In this case, the evidence has been evaluated to determine if the program is working.

In our example, the sample data is the 20 men hires out of 30. This gives us a sample proportion of $20/30 = 66.67\%$. The symbol for sample proportion is p ˆ (said “p hat” - the symbol above the p is supposed to be "^", but the online textbook program does not properly show it).

Evaluating evidence in statistics

To evaluate the evidence, we want to determine the probability of observing the evidence (or even better evidence against the assumption) assuming the assumption is true. Once we determine this probability, we need to determine if the event is unlikely or not unlikely. That is, we want to determine if it unlikely we could have observed the evidence, if the assumption is true. Or is it not unlikely that we observed the evidence, if the assumption is true. If it is unlikely to have observed the evidence, then most likely there is something wrong with the assumption and the claim is likely true. If it is not unlikely to have observed the evidence, then we can’t actually conclude that there is something wrong with the assumption and we cannot conclude that the claim is true.

To go back to the court case example, if you are a juror, you have to evaluate how unlikely or not unlikely it is that the defendant would have had a heated argument with the victim, and was found covered in blood and holding the murder weapon at the scene, if the defendant was innocent. If you think that it is unlikely that all of the pieces of evidence could have happened if the defendant is innocent, then you would find the defendant guilty. That is, the evidence calls into question the assumption. If you think that this it is not unlikely that all of these pieces of evidence could have happened if the defendant is innocent, then you would find the defendant not guilty. Notice that we don't conclude that the defendant is innocent. That is, we can't say that they are innocent; we can only say that they are not guilty.

When evaluating evidence, we are trying to evaluate the claim (i.e. not the position of skepticism). Therefore, the evidence has been collected about claim. No evidence has been collected about the assumption. Therefore, our conclusion can only be about the claim and not the assumption.

Therefore, if the probability is small and therefore unlikely, we can say that there is enough evidence to suggest that the assumption is likely false (i.e. guilty).

If the probability is not small and therefore not unlikely, we can say that there is not enough evidence to suggest that the assumption is false (i.e. not guilty).

In statistics, if the probability of an event happening is less than 1%, we say that the event is unlikely to happen. If the probability is greater than 10%, we say the event is not unlikely to happen. If the probability is between 1% and 10%, then it is up to the researcher to determine whether they believe that the event is unlikely or not unlikely. Usually, the researcher decides on the threshold between unlikely and not unlikely before performing the experiment or study.

In our example, to evaluate the evidence, we want to work out what is the probability this company would have hired 20 men out 30 (or even better evidence against the assumption) if the proportion of men hires is still 75%. That is, we want to find $P(X \leq 20)$, given $\pi = 75\%$). Notice that this is a conditional probability and the condition is the assumption.

What does “or even better evidence against the assumption” mean? It means that we don’t just find the probability of exactly 20 out of 30 men hires. We find the probability of at most 20 out of 30 men hires because if the company hired 19, 18, 17, … men then that would be even better evidence that 75% is no longer correct (as the sample proportion is getting more and more different from the assumed population proportion).

Why do we look at “or even better evidence against the assumption”? Often the probability of exactly one event happening is quite small. For example, the probability of getting exactly 10 heads out of 20 coin tosses is 17.62%, even though that is the most likely event to occur. Therefore, if we only looked at the probability of exactly one event happening (i.e. $P(X=20)$) rather than $P(X \leq 20)$, we may come to the false impression that an event is unlikely, when it could actually be explained by normal sampling variability.

Finding the probability

To find the probability, we need to find an appropriate distribution that models the situation. In later chapters, we will look at other models. Right now, the model we are going to use is the binomial distribution. For us to use this model, we have to ensure that the situation is meeting all of the conditions of the binomial distribution.

1. *The variable being studied is random*: This is not necessarily the case here as the applicants are not random and the hiring process is not random. If we randomly selected 30 hires from a greater number of hires, then it would be.
2. *The outcomes of the variable are being counted*: We are counting the number of men hired.
3. *There are a fixed number of trials*: We are looking at 30 hires ($n = 30$)
4. *There are only two possible outcomes*: Either the hire is a man or the hire is not a man.
5. *The n trials are independent and the probability of success and probability of failure remain constant*: This is true because we are assuming that the probability of hiring a man remains constant at 75%.

Though the first condition is not met, we can still use the binomial distribution to model the situation. That the model is not perfectly met would be a limitation of the study. That means that we would want to put a caveat at the end of our conclusion to state that this might reduce the accuracy of our results.

If the conditions of randomness and independence may not be fully met, then we can still utilize the binomial distribution. But we do have to be wary of the results. The other three conditions do need to be met to use the binomial distribution.

Now that we have the model, we can find the probability. In A computer program, we will use the binomial distribution with $n = 30$ and π (or probability of occurrence) $= 0.75$. Then we will find $P(X \leq 20)$.

From the computer program, we get $P(X \leq 20) = 0.19659 = 19.659$. Again, this probability is found under the assumption that the program has not worked (i.e. $\pi = 75\%$)

Evaluating the probability

The probability that we would have observed at most 20 hires that were men out of 30, under the assumption that the program did not work, is 19.659%. Therefore, it is *not unlikely* that we could have observed this evidence as the probability is greater than 10%. This means that having 20 out of 30 hires being men falls within the normal sampling variability for this data.

Based on the evidence collected, there is not sufficient evidence to suggest that the program worked. Notice we don't conclude that the program is not working.

In statistics, we never use the words “prove” or “true” when making a conclusion. All of our conclusions are based off of sample data that we are using to make a conclusion about the population. Therefore, there is always the chance of error.

Example

Olivier has spent five years honing his archery skills in various seedy locals around the world. Now he has returned to his city of birth to use these skills to take out criminals. One night while drinking vodka with his friends, he boasts that he can shoot an arrow into the bullseye, blindfolded at a distance of 50m 90% of the time.

“I don’t believe you!” Jack, Olivier’s best friend, slurred.

“I swear! I’ve really honed my skills.” Olivier countered.

“But remember last week when we were in that darkened factory, you missed two of your shots!" Thelma, Olivier's sister, countered.

"No. I meant to miss them."

Jack thought for a moment. "I think you are exaggerating and I'm going to test you."

"You're on!" Olivier sneered arrogantly.

To test that Olivier was exaggerating about his marksmanship, Jack set up a bunch of targets and, randomly had Olivier attempt the shot. Olivier hit the bullseye (blindfolded at a distance of 50m) 39 out of 50 times.

1. If Olivier's is not exaggerating, how many times out of 50 do we typically expect him to hit the bullseye? Write your answer as a range that takes into account variation.

Answer: We would expect Olivier to hit the bullseye 45 times give or take 2.121 times. This means a typical range is 42.88 to 47.12 bullseyes out of 50.

1. Based on your answer in a), is 39 out of 50 times potentially abnormal? Explain.

Answer: Since 39 is outside of the range, it would be deemed atypical, but that does not necessarily mean that it is abnormal.

1. What assumption do we need to make before determining whether the 39 out of 50 provides evidence for or against Olivier exaggerating?

Answer: Since Jack wants to show that Olivier is exaggerating, we want to assume that Olivier is not exaggerating. This means we want to assume $\pi = 90\%$, where π is the proportion of bullseyes that Olivier hits.

1. What model (i.e. distribution) will you use to test the evidence against the assumption? Explain why it is the best model to use. Note: This situation might not completely fit the model, but explain why it is still a reasonable model to use.

Answer: The distribution satisfies the conditions of the binomial distribution:

- *The variable being studied is random*: Since Jack is randomly having Olivier take the shot, we can say this is a random event.
- *The outcomes of the variable are being counted*: We are counting the number of bullseyes.
- *There are a fixed number of trials*: We are looking at 50 shots with the bow and arrow.
- *There are only two possible outcomes*: Either the shot is a bullseye or it is not.
- *The n trials are independent and the probability of success and probability of failure remain constant*:

This is true because we are assuming that the probability of hitting the bullseye remains constant at 90%.

1. What probability do you need to find to evaluate the evidence against the assumption?

Answer: We need to find the probability that Olivier hits at most 39 out of 50 bullseyes, assuming his accuracy is 90%. NOTE: We look at "at most 39" because having less bullseyes is even better evidence that Olivier is exaggerating (i.e. better evidence against the assumption).

1. Find that probability.

Answer: $P(X \leq 39 \text{ given } \pi = 90\%) = 0.00935 = 0.94\%$, (from computer program with $n = 50$, π (or probability of occurrence) $= 90\%$).

1. In the context of the problem, interpret the probability.

Answer: The probability that Olivier hit at most 39 out of 50 bullseyes, under the assumption that he wasn't exaggerating about his accuracy is 0.94%.

1. Does the probability provide evidence to support whether Olivier is exaggerating or not? Explain.

Answer: Since the probability that we observed our sample data is less than 1%, then it is unlikely that that Olivier is not exaggerating (i.e. that his accuracy is 90%). Therefore, it is likely that Olivier is exaggerating and cannot hit the bullseye 90% of the time blindfolded from 50m.

As stated in a previous question, the chance of an CRA audit for a tax return with over $25,000 in income is about 2% per year. An employee at I&S Square, a company that helps individuals do their yearly tax returns and helps if there is an audit, has noticed that people in Seba Beach, Alberta appear to have a greater chance of an audit than the rest of Canadians. Out of a random sample of 45 residents, four of them have been audited.

1. If the residents of Seba Beach are being audited fairly, how many residents out of 45 do we typically expect to get audited in a year? Write your answer as a range that takes into account variation.
2. Based on your answer in a), is 4 out of 45 audits potentially abnormal? Explain.
3. What assumption do we need to make before determining whether the 4 out of 45 audits is unfair?
4. What model (i.e. distribution) will you use to test the assumption? Explain why it is the best model to use. Note: This situation might not completely fit the model, but

explain why it is still a reasonable model to use.
5. What probability do you need to find to evaluate the assumption?
6. Find that probability.
7. In the context of the problem, interpret the probability.
8. Does the probability provide evidence to support or refute whether residents of Seba Beach are being unfairly audited? Explain.

---

1. We would expect 0.9 give or take 0.939 to be audited. This means a typical range is 0 to 1.8 residents to be audited out of 45.
2. Since 4 is outside of the range, it would be deemed atypical, but that does not necessarily mean that it is abnormal.
3. Since the employee at I&S Square wants to show that something strange is happening in Seba Beach, they would want to assume that nothing strange is happening. That is, the rate of audits is the same in Seba Beach as anywhere in Canada. This means we want to assume $\pi = 2\%$, where π is the proportion of people who are audited.
4. The binomial distribution:

- *The variable being studied is random*: We are looking at a random sample.
- *The outcomes of the variable are being*

counted: We are counting the number of audits.

- *There are a fixed number of trials*: We are looking at 45 residents.
- *There are only two possible outcomes*: Either the resident is audited or they are not.
- *The n trials are independent and the probability of success and probability of failure remain constant*: This is true because we are assuming that the probability of being audited remains constant at 2%.

1. We need to find the probability that at least 4 out of 45 residents are audited, assuming an audit rate of 2%.
2. P(X$\geq$4 given π = 2%) = 0.01242 = 1.24% (from computer program with n = 45, π (or probability of occurrence) = 2%).
3. The probability that at least 4 out of 45 Seba Beach residents are audited, under the assumption that the audit rate is 2%, is 1.24%.
4. Since the probability that we observed our sample data is between 1% and 10%, then we have to determine if the probability is unlikely or not unlikely. Since it is closer to 1% than 10%, we can say that the sample data is unlikely to have occurred under the assumption. Therefore, the evidence suggests that there is something

> wrong with the assumption. That is, there is evidence that the residents of Seba Beach are being audited at a higher rate than the rest of Canada.

Executives at Bull, a Canadian own cell phone company, are not very happy with their current customer satisfaction surveys. Using a Likert scale, they surveyed a very large sample of clients who phoned Bull and spoke to a customer service representative. They have determined that only 60% of customers rate their overall satisfaction with the service they received at 4 or higher. That is, they either strongly agree or agreed with the statement, “I am happy with the overall customer service I received during my most recent call to Bull.”

They feel that this is too low as 40% of customers were not happy with their service. To address these issues, they’ve brought in a consultant who has suggested that customers are happier with their service if they feel they’ve built a rapport with the customer service representative. Thus, Bull has decided to train their customer service representatives to start each call with a short conversation. As customers are from across Canada and it would be bad if the conversations were generic, to help their customer service representatives build rapport, a short notice shows up on their screens before they take the call that contains suggested conversation topic for the area the person is calling from. For example, it might include information about weather in the local area and how the local sporting team has done in their most recent game.

After the customer service representatives have been trained in how to make small talk to build rapport, a random sample of sixty customers who called Bull and spoke to a customer service representative is taken. The participants are asked the same question about their overall satisfaction with their customer service phone call as stated above. The results of the survey are listed below:

1	1	1	1	1	1	2	2	2	2	2	3	3	3	3	3	3	3	3	4
4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4
4	4	4	4	4	4	4	4	4	4	5	5	5	5	5	5	5	5	5	5

Does the recent sample provide sufficient evidence to suggest that the proportion of customers who are happy with their overall service when they call Bull has increased from

60%? Explain your answer in detail.

To be skeptical, we want to assume that the program has not worked (i.e. π stayed at 60%). The evidence is 41 out of 60 customers gave a rating of 4 or 5. Perfect evidence that the program worked would be 60 out of 60 happy customers in every single sample. The probability that we would observe at least 41 out of 60 customers who gave a score of four or five regarding their overall satisfaction, assuming that the new program has not worked, is 11.70% (from a computer program with $n = 60$, $\pi = 0.60$). Therefore, it is not unlikely that we observed the evidence that we did, under the assumption the program did not work. This means that we cannot conclude that the program worked.

What we can conclude when the probability is “not unlikely”: If the probability is greater than 10%, then it means that it is not unlikely that we observed this evidence under the assumption. We can NOT conclude that the assumption is likely true as the evidence was collected to evaluation the claim (not the assumption). Instead, we can only conclude that there is not enough evidence to say that the claim is true. When the probability is “not unlikely”, we have really learned very little about the claim.

Chapter Review

A statistical experiment can be classified as a binomial experiment if the following conditions are met:

1. The variable being studied is random.
2. The outcomes of the variable are being counted.
3. There are a fixed number of trials. The letter *n* denotes the number of trials.
4. There are only two possible outcomes, called "success" and "failure," for each trial π denotes the probability of a success on one trial, and $1-\pi$ denotes the probability of a failure on one trial.
5. The *n* trials are independent and are repeated using identical conditions. Because the *n* trials are independent, the outcome of one trial does not help in predicting the outcome of another trial. Another way of saying this is that for each individual trial, the probability, π , of a success and probability, $1- \pi$, of a failure remain constant.

The outcomes of a binomial experiment fit a binomial probability distribution. The random variable $X =$ the number of successes obtained in the n independent trials. The mean of X can be calculated using the formula $\mu = n\pi$, and the standard deviation is given by the formula $\sigma = n\pi$ $(1-\pi)$.

To evaluate evidence, we must first begin from a position of skepticism (i.e. assume the opposite of what we want to show). Then we must find a probability which is the distance from the actual evidence to perfect evidence against the assumption. We can then evaluate the probability by determining whether it is less than 1% (which means it is unlikely the evidence occurred under the assumption) or if it is greater than 10% (which means it is not unlikely the evidence occurred under the assumption). If the probability is deemed unlikely, then we reject the assumption, which means there is enough evidence to support what we originally wanted to show (the claim). If the probability is deemed not unlikely, then we do not reject the assumption, which means there is not enough evidence to support what we originally wanted to show (the claim). In the latter situation, we cannot make any conclusions about the assumption as the evidence was collected only for the claim.

Practice

The first few exercises provided are from the textbook Business Statistics -- BSTA 200 -- Humber College -- Version 2016RevA -- DRAFT 2016-04-04 by Alexander Holmes, Lyryx Learning: http://cnx.org/contents/f3aefa9e-58d2-41ea-969f-04dc2cb04c82@5.20

Use the following information to answer the next seven exercises: The Higher Education Research Institute at UCLA collected data from 203,967 incoming first-time, full-time freshmen from 270 four-year colleges and universities in the U.S. 71.3% of those students replied that, yes, they believe that same-sex couples should have the right to legal marital status. Suppose that you randomly pick eight first-time, full-time freshmen from the survey. You are interested in the number that believes that same sex-couples should have the right to legal marital status.

In words, define the random variable *X*.

$X =$ the number that reply "yes"

What values does the random variable *X* take on?

0, 1, 2, 3, 4, 5, 6, 7, 8

Construct the probability distribution function (PDF). That is, fill in the table below. In the left column put in the possible values for X. In the right column, put in the probability for exactly X, i.e. P(X=x)

x	P(X=x)

x	P(X=x)
0	0.00005
1	0.0009
2	0.0080
3	0.0395
4	0.1227
5	0.2439
6	0.3030
7	0.2151
8	0.0668

On average (μ), how many would you expect to answer yes?

5.7

What is the standard deviation (σ)?

1.2795

What is the probability that at most five of the freshmen reply “yes”?

0.4151

What is the probability that at least two of the freshmen reply “yes”?

0.9990

A school newspaper reporter decides to randomly survey 12 students to see if they will attend Tet (Vietnamese New Year) festivities this year. Based on past years, she knows that

18% of students attend Tet festivities. We are interested in the number of students who will attend the festivities.

1. In words, define the random variable X.
2. List the values that X may take on.
3. How many of the 12 students do we expect to attend the festivities?
4. Find the probability that at most four students will attend.
5. Find the probability that more than two students will attend.

---

1. $X =$ the number of students who will attend Tet.
2. 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12
3. 2.16
4. 0.9511
5. 0.3702

Use the following information to answer the next three multiple choice questions: The probability that the Calgary Flames will win any given game is 0.4617 based on a 45-year win history of 1,616 wins out of 3,500 games played (as of Sept. 2017). An upcoming monthly schedule contains 12 games.

The expected number of wins for that upcoming month is:

1. 1.67
2. 12
3. 1616 3500
4. 5.54

---

d. 5.54

Let $X =$ the number of games won in that upcoming month.

What is the probability that the Calgary Flames win exactly six games in that upcoming month?

1. 0.2178
2. 0.4167
3. 0.7664
4. 0.7116

---

a

What is the probability that the Calgary Flames win at least five games in that upcoming month

1. 0.2176
2. 0.2762
3. 0.7238
4. 0.5062

---

c

The chance of an Canadian Revenue Agency audit for a tax return with over $25,000 in income is about 2% per year. We are interested in the expected number of audits a person with that income has in a 20-year period. Assume each year is independent.

1. In words, define the random variable X.
2. List the values that X may take on.
3. How many audits are expected in a 20-year period?
4. Find the probability that a person is not audited at all.
5. Find the probability that a person is audited more than twice.

1. $X =$ the number of audits in a 20-year period
2. 0, 1, 2, …, 20
3. 0.4
4. 0.6676
5. 0.0071

According to The World Bank, only 9% of the population of Uganda had access to electricity as of 2009. Suppose we randomly sample 150 people in Uganda. Let $X =$ the number of people who have access to electricity.

1. Calculate the mean and standard deviation of X.
2. Find the probability that 15 people in the sample have access to electricity.
3. Find the probability that at most ten people in the sample have access to electricity.
4. Find the probability that more than 25 people in the sample have access to electricity.

---

1. Mean $= np = 150(0.09) = 13.5$; Standard Deviation $=$ npq $=$ $150(0.09)(0.91) \approx 3.5050$
2. $P(x = 15) = 0.0988$
3. $P(x \leq 10) = 0.1987$
4. $P(x > 25) = 0.0009$

Jenna and Megan looked at the new packaging.

“I guess it looks ok.” Megan hedged.

“The design team says that this new packaging really sells the time-saving nature of the kit.”

“But it’s kinda off-putting.” They continued to stare at the new packaging. Jenna and Megan had developed a make-up kit called ‘5 minute make-up’, which was aimed at women on the go who wanted to put ‘their face on’ but a lot quicker than they usually did. Their target market was new moms, moms with full-time jobs, full-time students working full-time, …. In other words, anyone who didn’t have 30 minutes every morning to do their make-up. Their little start-up was doing well. They’d arranged for their product to be produced, had made how-to videos on YouTube, and were starting to get their products put into stores. Their dream placement was in Sephora.

Now that they were more established, they had decided to hire a marketing expert who could help take them to the next level. The first thing that Leticia suggested was to change the packaging. She argued that their old packaging didn’t convey the premise of the product clearly enough. With the help of a design team, Leticia had come up with a packaging that showed a harried woman with hair everywhere and bags under her eyes looking overwhelmed. But when you flip the package over, the woman was now perfectly put together – ‘only five-minute make-up can save you from being a hot mess’.

Jenna finally broke the silence. “I don't know if I would even pick up this package. It just looks so depressing. But what do we do?”

“Leticia wants to put the product in this new packaging in five stores that carry our products. Based off of previous sales numbers, we know that the stores sell 68% of the product we give them in a two-week period.”

“How does that help us? Do we just watch our sales plummet?” Jenna was sounding exasperated.

“I’m getting to that.” Megan soothed. “Leticia is convinced this packaging will increase sales. But what if we can show her that it doesn’t? Let’s put this packaging into the five stores and then see how many kits were actually sold. I bet that we can show her that the sales went down.”

“I don’t see how that is useful. We still have to pay her stupid fee.”

“You should read her contract more closely. She only gets paid if she can show that sales increased. If they don’t, then not only does she not get paid but she also has to pay for any contractors (i.e. the design team).”

Jenna perked up visibly at this.

Over the next two weeks, five stores carried the new packaging. Megan and Jenna provided each store with 100 kits. At the end of the two weeks, 306 of the kits were sold.

1. What is the observation unit? What is the variable? Categorize it.
2. What do Jenna and Megan want to show?
3. What assumption do Jenna and Megan need to make in order to investigate your answer in question 2? Write your answer both as a sentence and as a probability.
4. What is the evidence that Jenna and Megan have found?
5. Describe the process that Jenna and Megan will go through to evaluate this evidence. Your description should include (but is not limited to) what probability they will find and what they will do with that probability once they've found it. Don't actually do the process (that comes later). Just describe what they will do.
6. Jenna and Megan believe that the binomial distribution will be the best model to find the required probability. Does this situation meet the criteria for a binomial? Examine each criterion and comment on whether it is satisfied here or not.
7. Regardless of your answer above, use a binomial distribution to model this situation. Find the appropriate probability to evaluate the evidence using MegaStat.

8. In sentence form, explain what the probability you have found means in the context of the question. Do not make a conclusion yet. Instead explain what it is a probability of.
9. Now make a conclusion. In particular, answer this question: Is their enough evidence to suggest that Leticia’s new packaging has reduced sales? Justify your answer.

---

1. Observational unit: Five-minute make-up kit; Variable: Did it sell or not; Categorize: Categorical
2. They want to show that the new packaging will decrease sales.
3. They need to assume the opposite of what they want to show. Therefore, they need to assume that the new packaging does not decrease sales. Therefore, the proportion of kits sold stays the same at 68%.
4. They have found that out of 500 kits supplied, 306 of them have been sold.
5. They first need to start with an assumption (= 68%). Then they need to come up with a model based on this assumption. Once they have the model, they will use it to find the probability that the stores sold at most 306 out of 500 kits, assuming that the new packaging has not decreased sales

(i.e. stayed at 68%). Once they have the probability, they need to determine whether the event is likely or unlikely. An event is unlikely if the probability is less than 1%. An event is likely if the probability is more than 10%. If the event is unlikely, then it means that it is unlikely we observed the evidence under the assumption. Since we know the evidence actually happened, that makes us question the assumption. Thus, it is unlikely the assumption is true based off of the evidence. If the event is likely to happen, then the assumption is likely to be true based off of the evidence.

6.
   - Is the data randomly collected? Most likely not. The 500 kits that we are looking at were not randomly selected.
   - Is the data discrete (countable)? As we are counting the number of kits that are sold the data is discrete.
   - Are the events independent? This may be a fair assumption for this study. Most likely the sale of one kit is not dependent on whether another kit is sold. Though if two friends buy the kits together or someone buys a bunch as presents, this is not the case, but in general it is more likely independent

than dependent.

- Are there a fixed number of trials? In this case, the number of trials would be the 500 kits with the new packaging.
- Are there two possible outcomes? Either a kit is sold or it is not.

12. $P(X \leq 36) = 0.00077 = 0.077\%$
13. The probability that we observed at most 306 out of 500 kits sold (1), assuming the rate of sales is 68%, is 0.077%
14. Since the probability is less than 1%, then it is very unlikely that we would have observed this evidence under the assumption. Since we actually observed the evidence but assumed that the rate was 68%, what we have assumed is called into question. Therefore, it is unlikely that the assumption is true . Therefore, it is likely that the new packaging has resulted in a decrease in sales .

Striking Donkey Coffee recently sold an 80% stake in their company to Baravalle, an Italian coffee conglomerate. Striking Donkey's logo is simplistic. Baravalle wants to maintain brand recognition, but also wants to put their stamp on the company. In particular, Baravalle is known for its modern and stylish advertisements.

Designers and marketers at Baravalle have worked tirelessly for the last month to come up with two revised Striking Donkey logos (not included, because it is top secret). They are referred to as Logo 1 and Logo 2.

Now they want to determine whether customers show any preference to either logo. To do this, they asked a random sample of 40 customers who were familiar with the Striking Donkey Brand which logo they prefer. Participants had to make a choice between the logos.

The results of the study were that 26 out of the 40 participants preferred Logo 2.

The marketers at Baravalle now want to do a statistical analysis to determine whether Logo 2 is preferred significantly more than Logo 1.

1. What assumption do you need to start with when determining whether Logo 2 is preferred significantly more than Logo 1? State your answer both in a sentence and mathematically.
2. Can this situation be modelled by the binomial distribution? Support your answer by showing why or why not this situation satisfies each of the five criteria of the binomial distribution.

3. After previous issues with horrible new logo launches, Baravalle only wants to go forward if there is clear evidence that Logo 2 is preferred. Based on this, what level of significance should they use? Explain your reasoning.
4. Regardless of your answer in b, assume that this situation satisfies the binomial distribution for the remainder of the question. Use a computer program to find that appropriate probability that will allow you to evaluate the evidence.
5. In sentence form, explain what the probability you have found means in the context of the question. Do not make a conclusion yet. Instead explain what it is a probability of.
6. Based on the probability, determine whether Logo 2 is preferred significantly more than Logo 1. Explain your reasoning.

Observational unit: People who are familiar with Striking Donkey Coffee; Variable: Whether they prefer Logo 2; Type of variable: Categorical.

1. They need to assume the opposite of what they want to show. This means they need to assume that Logo 2 is NOT preferred significantly more than Logo 1. This would

mean they are preferred equally. Therefore, there is a 50% chance that someone will choose Logo 2.

2.
   - The data is collected randomly: Yes. It is a random sample of participants.
   - The outcomes are counted: Yes. They count how many people like Logo 2.
   - There are two possible outcomes: Yes. Either they prefer Logo 2 or they did not.
   - There are a fixed number of trials: Yes. They asked 40 people.
   - The trials are independent of each other. Yes. It is fair to assume that no participant's preference is based on another participants preference.

8. The more unlikely it is that we observed our evidence, the smaller the probability will be. This means, the smaller the probability, the more unlikely it is that the assumption (i.e. that there is no preference between the logos) is true. Since the marketers want clear evidence that there is a preference, they want a smaller probability, which would show it is unlikely that there is a preference between the logos. The level of significance is the threshold between likely and unlikely. Thus, if they want clear evidence, they

want to set their threshold “high”, meaning they want to make it a small number. Since the level of significance is between 1% and 10%, the lowest level of significance (meaning the highest threshold of evidence) is at 1%.

9. $P(X \geq 26) = 0.04035 = 4.04\%$
10. The probability that we observed at least 26 out of 40 people who preferred Logo 2, assuming that there was no preference between the logos, is 4.04%.
11. Since the probability is greater than 1% (it is 4.04%), it is not unlikely that we observed at least 26 out of 40 people who preferred Logo 2, assuming that there was no preference between the logos. Therefore, we do not reject that there was no preference between the logos. This suggests that Logo 2 is NOT preferred significantly more than Logo 1.

Introduction to the Normal Distribution

class="introduction" If you ask enough people about their shoe size, you will find that your graphed data is shaped like a bell curve and can be described as normally distributed. (credit: Ömer Ünlü)

Chapter Objective

By the end of this chapter, the student should be able to:

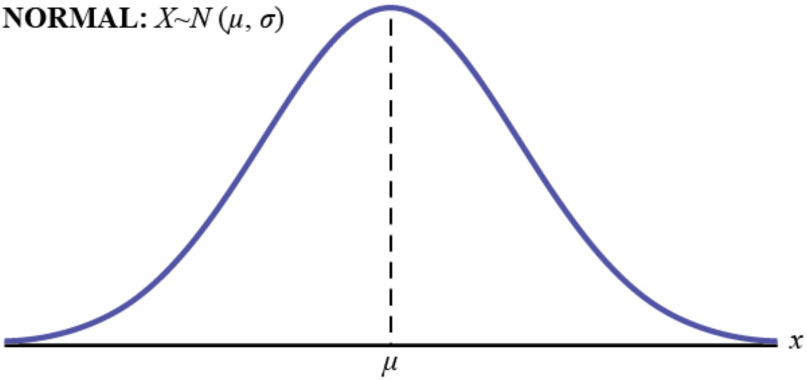
- Recognize the normal probability distribution and apply it appropriately.
- Recognize the standard normal probability distribution and apply it appropriately.
- Compare normal probabilities by converting to the standard normal distribution.

The normal probability density function, a continuous distribution, is the most important of all the distributions. It is widely used and even more widely abused. Its graph is bell-shaped. You see the bell curve in almost all disciplines. Some of these include psychology, business, economics, the sciences, nursing, and, of course, mathematics. Some of your instructors may use the normal distribution to help determine your grade. Most IQ scores are normally distributed. Often real-estate prices fit a normal distribution.

The normal distribution is extremely important, but it cannot be applied to everything in the real world. Remember here that we are still talking about the distribution of population data. This is a discussion of probability and thus it is the population data that may be normally distributed, and if it is, then this is how we can find probabilities of specific events just as we did for population data that may be binomially distributed or Poisson distributed. This caution is here because in the next chapter we will see that the normal distribution describes something very different from raw data and forms the foundation of inferential statistics.

In this chapter, you will study the normal distribution, the standard normal distribution, and applications associated with them.

The normal distribution has two parameters (two numerical descriptive measures), the mean (μ) and the standard deviation (σ). If X is a quantity to be measured that has a normal distribution with mean (μ) and standard deviation (σ), we designate this by writing the following formula of the normal probability density function:



The curve is symmetrical about a vertical line drawn through the mean, μ. The mean is the same as the median, which is the same as the mode, because the graph is symmetric about μ. As the notation indicates, the normal distribution depends only on the mean and the standard deviation. Note that this is unlike several probability density functions we have already studied, such as the Poisson, where the mean is equal to µ and the standard deviation simply the square root of the mean, or the binomial, where p is used to determine both the mean and standard deviation. Since the area under the curve must equal one, a change in the standard deviation, σ, causes a change in the shape of the curve; the curve becomes fatter and wider or skinnier and taller depending on σ. A change in μ causes the graph to shift to the left or right. This means there are an infinite number of normal probability distributions. One of special interest is called the **standard normal distribution**.

Formula Review

$X \sim N(\mu, \sigma)$

μ = the mean σ = the standard deviation

Glossary

Normal Distribution

a continuous random variable (RV) with pdf $f(x)$ = 1 σ 2π e –(x – μ) 2 σ 2 2 , where μ is the mean of the distribution and σ is the standard deviation; notation: $X \sim N(\mu, \sigma)$. If $\mu = 0$ and $\sigma = 1$, the RV, Z, is called the **standard normal distribution**.

The Standard Normal Distribution

The **standard normal distribution** is a normal distribution of **standardized values called *z*-scores**. **A *z*-score is measured in units of the standard deviation.** For example, if the mean of a normal distribution is five and the standard deviation is two, the value x = 11 is three standard deviations above (or to the right of) the mean. The calculation is as follows:

$$x = \mu + (z)(\sigma) = 5 + (3)(2) = 11$$

The *z*-score is three.

The mean for the standard normal distribution is zero, and the standard deviation is one. What this does is dramatically simplify the mathematical calculation of probabilities. Take a moment and substitute zero and one in the appropriate places in the above formula and you can see that the equation collapses into one that can be much more easily solved using integral calculus. The transformation z = x−μ σ produces the distribution $Z \sim N(0, 1)$. The value x comes from a known normal distribution with known mean μ and known standard deviation σ. The *z*-score tells how many standard deviations a particular x is away from the mean.

Z-Scores

If *X* is a normally distributed random variable and *X* ~ *N(μ, σ)*, then the *z*-score is:

$z = x - \mu \ \sigma$

The *z*-score tells you how many standard deviations the value *x* is above (to the right of) or below (to the left of) the mean, *μ*. Values of *x* that are larger than the mean have positive *z*-scores, and values of *x* that are smaller than the mean have negative *z*-scores. If *x* equals the mean, then *x* has a *z*-score of zero.

Suppose *X* ~ *N(5, 6)*. This says that *x* is a normally distributed random variable with mean *μ* = 5 and standard deviation *σ* = 6. Suppose *x* = 17. Then:

$z = x – \mu \ \sigma = 17 – 5 \ 6 = 2$

This means that *x* = 17 is **two standard deviations** (2*σ*) above or to the right of the mean *μ* = 5. The standard deviation is *σ* = 6.

Now suppose *x* = 1. Then: *z* = $x – \mu \ \sigma$ = $1 – 5 \ 6$ = –0.67 (rounded to two decimal places)

This means that *x* = 1 is 0.67 standard deviations (–0.67*σ*) below or to the left of the mean *μ* = 5.

Some doctors believe that a person can lose five pounds, on average, in a month by reducing his or her fat intake and by exercising consistently. Suppose weight loss has a normal distribution. Let $X =$ the amount of weight lost(in pounds) by a person in a month. Use a standard deviation of two pounds. $X \sim N(5, 2)$.

Suppose a person **gained** three pounds (a negative weight loss). Then $z =$ _______. This z-score tells you that $x = –3$ is ______ standard deviations to the _______ (right or left) of the mean.

Z=x-μσ=-3-52=-4

$z = –4$. This z-score tells you that $x = –3$ is **four** standard deviations to the **left** of the mean.

Suppose the random variables X and Y have the following normal distributions: $X \sim N(5, 6)$ and $Y \sim N(2, 1)$. If $x = 17$, then $z = 2$. (This was previously shown.) If $y = 4$, what is z?

z = y−μ σ = 4−2 1 = 2 where $\mu = 2$ and $\sigma = 1$.

The z-score for $y = 4$ is $z = 2$. This means that four is $z = 2$ standard deviations to the right of the mean. Therefore, $x = 17$ and $y = 4$ are both two (of **their own**) standard deviations to the right of **their** respective means.

The *z*-score allows us to compare data that are scaled differently. To understand the concept, suppose $X \sim N(5, 6)$ represents weight gains for one group of people who are trying to gain weight in a six week period and $Y \sim N(2, 1)$ measures the same weight gain for a second group of people. A negative weight gain would be a weight loss. Since $x = 17$ and $y = 4$ are each two standard deviations to the right of their means, they represent the same, standardized weight gain **relative to their means**.

Try It

Fill in the blanks.

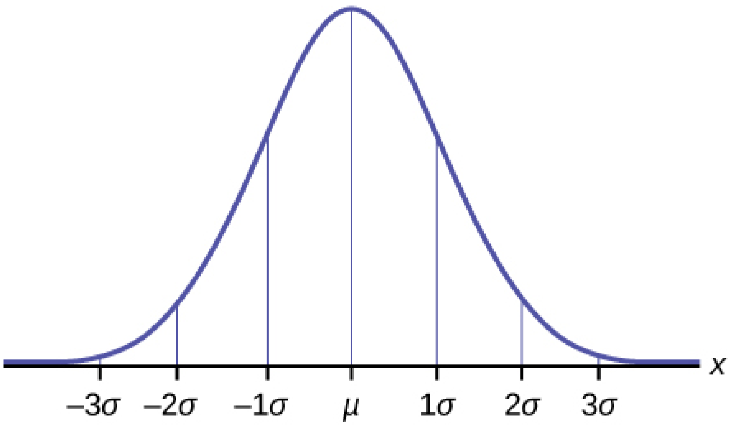
Jerome averages 16 points a game with a standard deviation of four points. $X \sim N(16,4)$. Suppose Jerome scores ten points in a game. The *z*–score when $x = 10$ is –1.5. This score tells you that $x = 10$ is ____ standard deviations to the ____(right or left) of the mean____(What is the mean?).

1.5, left, 16

The Empirical Rule
If X is a random variable and has a normal distribution with mean μ and standard deviation σ, then the **Empirical Rule** says the following:

- About 68.26% of the x values lie between -1σ and $+1\sigma$ of the mean μ (within one standard deviation of the mean).
- About 95.44% of the x values lie between -2σ and $+2\sigma$ of the mean μ (within two standard deviations of the mean).
- About 99.73% of the x values lie between -3σ and $+3\sigma$ of the mean μ (within three standard deviations of the mean). Notice that almost all the x values lie within three standard deviations of the mean.
- The z-scores for $+1\sigma$ and -1σ are $+1$ and -1, respectively.
- The z-scores for $+2\sigma$ and -2σ are $+2$ and -2, respectively.
- The z-scores for $+3\sigma$ and -3σ are $+3$ and -3 respectively.

The empirical rule is also known as the 68-95-99.7 rule.



The mean height of 15 to 18-year-old males from Chile from 2009 to 2010 was 170 cm with a standard deviation of 6.28 cm. Male heights are known to follow a normal distribution. Let X = the height of a 15 to 18-year-old male from Chile in 2009 to 2010. Then $X \sim N(170, 6.28)$.

a. Suppose a 15 to 18-year-old male from Chile was 168 cm tall from 2009 to 2010. The z-score when x = 168 cm is z = _____. This z-score tells you that x = 168 is ______ standard deviations to the ______ (right or left) of the mean ____ (What is the mean?).

$Z = x\text{-}\mu\sigma = 168\text{-}1706.28 = \text{-}0.32$

a. –0.32, 0.32, left, 170

b. Suppose that the height of a 15 to 18-year-old male from Chile from 2009 to 2010 has a *z*-score of $z = 1.27$. What is the male's height? The *z*-score ($z = 1.27$) tells you that the male's height is ______ standard deviations to the _______ (right or left) of the mean.

Z=x-
μσ=x-1706.28=1.27→1.27*6.28+170=177.98

b. 177.98, 1.27, right

Try It

In 2012, 1,664,479 students took the SAT exam. The distribution of scores in the verbal section of the SAT had a mean $\mu = 496$ and a standard deviation $\sigma = 114$. Let X = a SAT exam verbal section score in 2012. Then $X \sim N(496, 114)$.

Find the *z*-scores for $x_1 = 325$ and $x_2 = 366.21$. Interpret each *z*-score. What can you say about $x_1 = 325$ and $x_2 = 366.21$?

The z-score for $x_1 = 325$ is $z_1 = –1.14$.

The z-score for $x_2 = 366.21$ is $z_2 = –1.14$.

Student 2 scored closer to the mean than Student 1 and, since they both had negative z-scores, Student 2 had the better score.

Suppose x has a normal distribution with mean 50 and standard deviation 6.

- About 68% of the x values lie between $–1\sigma = (–1)(6) = –6$ and $1\sigma = (1)(6) = 6$ of the mean 50. The values $50 – 6 = 44$ and $50 + 6 = 56$ are within one standard deviation of the mean 50. The z-scores are –1 and +1 for 44 and 56, respectively.
- About 95% of the x values lie between $–2\sigma = (–2)(6) = –12$ and $2\sigma = (2)(6) = 12$. The values $50 – 12 = 38$ and $50 + 12 = 62$ are within two standard deviations of the mean 50. The z-scores are –2 and +2 for 38 and 62, respectively.
- About 99.7% of the x values lie between $–3\sigma = (–3)(6) = –18$ and $3\sigma = (3)(6) = 18$ of the mean 50. The values $50 – 18 = 32$ and $50 + 18 = 68$ are within three standard deviations of the mean 50. The z-scores are –3 and +3 for 32 and 68, respectively.

Try It

Suppose X has a normal distribution with mean 25 and standard deviation five. Between what values of x do 68% of the values lie?

between 20 and 30.

Try It

The scores on a college entrance exam have an approximate normal distribution with mean, $\mu = 52$ points and a standard deviation, $\sigma = 11$ points.

1. About 68% of the y values lie between what two values? These values are ____________. The z-scores are ____________, respectively.
2. About 95% of the y values lie between what two values? These values are ____________. The z-scores are ____________, respectively.
3. About 99.7% of the y values lie between what two values? These values are ____________. The z-scores are ____________, respectively.

1. About 68% of the values lie between the values 41 and 63. The *z*-scores are –1 and 1, respectively.
2. About 95% of the values lie between the values 30 and 74. The *z*-scores are –2 and 2, respectively.
3. About 99.7% of the values lie between the values 19 and 85. The *z*-scores are –3 and 3, respectively.

References

"Blood Pressure of Males and Females." StatCruch, 2013. Available online at http://www.statcrunch.com/5.0/viewreport.php?reportid=11960 (accessed May 14, 2013).

"The Use of Epidemiological Tools in Conflict-affected populations: Open-access educational resources for policy-makers: Calculation of z-scores." London School of Hygiene and Tropical Medicine, 2009. Available online at http://conflict.lshtm.ac.uk/page_125.htm (accessed May 14, 2013).

“2012 College-Bound Seniors Total Group Profile Report.” CollegeBoard, 2012. Available online at http://media.collegeboard.com/digitalServices/pdf/research/TotalGroup-2012.pdf (accessed May 14, 2013).

“Digest of Education Statistics: ACT score average and standard deviations by sex and race/ethnicity and percentage of ACT test takers, by selected composite score ranges and planned fields of study: Selected years, 1995 through 2009.” National Center for Education Statistics. Available online at http://nces.ed.gov/programs/digest/d09/tables/dt09_147.asp (accessed May 14, 2013).

Data from the *San Jose Mercury News*.

Data from *The World Almanac and Book of Facts*.

“List of stadiums by capacity.” Wikipedia. Available online at https://en.wikipedia.org/wiki/List_of_stadiums_by_capacity (accessed May 14, 2013).

Data from the National Basketball Association. Available online at www.nba.com (accessed May 14, 2013).

Chapter Review

A z-score is a standardized value. Its distribution is the standard normal, $Z \sim N(0, 1)$. The mean of the z-scores is zero and the standard deviation is one. If z is the z-score for a value x from the normal distribution $N(\mu, \sigma)$ then z tells you how many standard deviations x is above (greater than) or below (less than) μ.

Practice Questions

In a normal distribution, $x = 3$ and $z = 0.67$. This tells you that $x = 3$ is ___ standard deviations to the ___ (right or left) of the mean.

0.67, right

In a normal distribution, $x = –5$ and $z = –3.14$. This tells you that $x = –5$ is ___ standard deviations to the ___ (right or left) of the mean.

3.14, left

About what percent of x values from a normal distribution lie within one standard deviation (left and right) of the mean of that distribution?

about 68%

About what percent of the *x* values from a normal distribution lie within two standard deviations (left and right) of the mean of that distribution?

about 95.44%

About what percent of *x* values lie between the second and third standard deviations (both sides)?

about 4%

Use the following information to answer the next two multiple choice exercises: The patient recovery time from a particular surgical procedure is normally distributed with a mean of 5.3 days and a standard deviation of 2.1 days.

What is the median recovery time?

1. 2.7
2. 5.3
3. 7.4
4. 2.1

b

What is the *z*-score for a patient who takes ten days to recover?

1. 1.5
2. 0.2
3. 2.2
4. 7.3

c

Wesley Crusher was tasked with exploring the **Selcundi Drema** sector. He found a new species of tribbles. In his final report, he stated, “Though tribbles vary in size and dimension, the middle 99.73% of them weigh between 4 and 7.2 kg and follow a normal distribution.” Based on this, what is the mean and standard deviation for the weight of tribbles? Choose the best answer.

1. mean = 5.6 kg, standard deviation = 1.07 kg
2. mean = 5.6 kg, standard deviation = 0.53 kg
3. mean = 5.6 kg, standard deviation = 0.8 kg

4. mean = 99.73 kg, standard deviation = 3.2 kg
5. There is not enough information to determine this.

---

b

Glossary

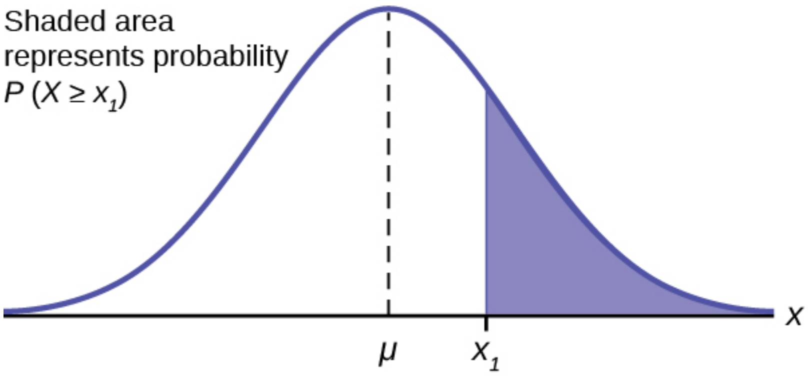
Standard Normal Distribution

a continuous random variable (RV) $X \sim N(0, 1)$; when X follows the standard normal distribution, it is often noted as $Z \sim N(0, 1)$.

z-score

the linear transformation of the form z = x-μ σ or written as z = |x – μ| σ ; if this transformation is applied to any normal distribution $X \sim N(\mu, \sigma)$ the result is the standard normal distribution $Z \sim N(0,1)$. If this transformation is applied to any specific value x of the RV with mean μ and standard deviation σ, the result is called the z-score of x. The z-score allows us to compare data that are normally distributed but scaled differently. A z-score is the number of standard deviations a particular x is away from its mean value.

The shaded area in the following graph indicates the area to the right of x. This area is represented by the probability $P(X > x)$. Normal tables, computers, and calculators provide or calculate the probability $P(X > x)$.

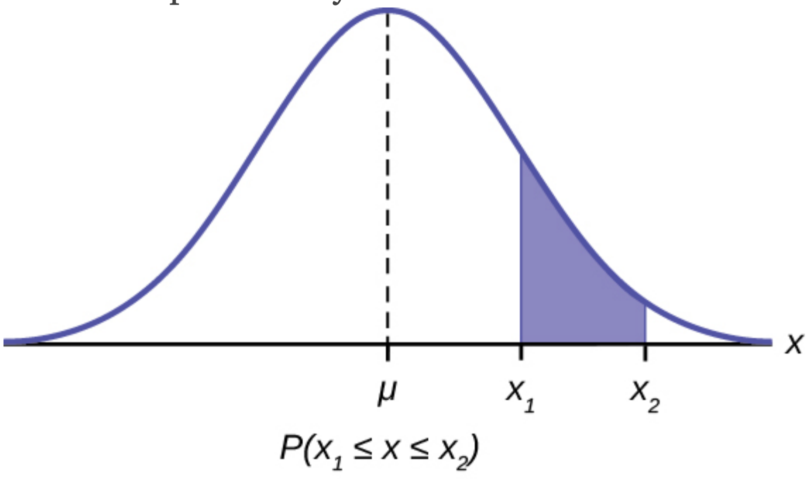


The area to the right is then $P(X > x) = 1 – P(X < x)$. Remember, $P(X < x) =$ **Area to the left** of the vertical line through x. $P(X < x) = 1 – P(X < x) =$ **Area to the right** of the vertical line through x. $P(X < x)$ is the same as $P(X \leq x)$ and $P(X > x)$ is the same as $P(X \geq x)$ for continuous distributions.

Calculations of Probabilities

To find the probability for probability curves with a continuous random variable we need to calculate the area under the curve across the values of X we are interested in. For the normal distribution this seems a difficult task given the complexity of the formula. There is, however, a simply way to get what we want.

We start knowing that the area under a probability curve is the probability.



This shows that the area between X1 and X2 is the probability as stated in the formula: P (X1 ≤ x ≤ X2)

The mathematical tool needed to find the area under a curve is integral calculus. The integral of the normal probability density function between the two points x1 and x2 is the area under the curve between these two points and is the probability between these two points.

Doing these integrals is no fun and can be very time consuming. But now, remembering that there are an infinite number of normal distributions out there, we can consider the one with a mean of zero and a standard deviation of 1. This particular normal distribution is given the name Standard Normal Distribution. Putting these values into the formula it reduces to a very simple equation. We can now quite easily calculate all probabilities for any value of x, for this particular normal distribution, that has a mean of zero and a standard deviation of 1. These have been produced and are available here in the text or everywhere on the web. They are presented in various ways. The table in this text is the most common presentation and is set up with probabilities for one-half the distribution beginning with zero, the mean, and moving outward. The shaded area in the graph at the top of the table represents the probability from zero to the specific Z value noted on the horizontal axis, Z.

The only problem is that even with this table, it would be a ridiculous coincidence that our data had a mean of zero and a standard deviation of one. The solution is to convert the distribution we have with its mean and standard deviation to this new Standard Normal Distribution. The Standard Normal has a random variable called Z.

Using the standard normal table, typically called the normal table, to find the probability of one standard deviation, go to the Z column, reading down to 1.0 and then read at column 0. That number, 0.3413 is the probability from zero to 1 standard deviation. At the top of the table is the shaded area in the distribution which is the probability for one standard deviation. The table has solved our integral calculus problem. But only if our data has a mean of zero and a standard deviation of 1.

However, the essential point here is, the probability for one standard deviation on one normal distribution is the same on every normal distribution. If the population data set has a mean of 10 and a standard deviation of 5 then the probability from 10 to 15, one standard deviation, is the same as from zero to 1, one standard deviation on the standard normal distribution. To compute probabilities, areas, for any normal distribution, we need only to convert the particular normal distribution to the standard normal distribution and look up the answer in the tables. As review, here again is the **standardizing formula**:

$Z=x-\mu\sigma$

where Z is the value on the standard normal distribution, X is the value from a normal distribution one wishes to convert to the standard normal, μ and σ are, respectively, the mean and standard deviation of that population. Note that the equation uses μ and σ which denotes population parameters. This is still dealing with probability so we always are dealing with the population, with **known** parameter values and a **known** distribution.

It is also important to note that because the normal distribution is symmetrical it does not matter if the z-score is positive or negative when calculating a probability. One standard deviation to the left (negative Z-score) covers the same area as one standard deviation to the right (positive Z-score). This fact is why the Standard Normal tables do not provide areas for the left side of the distribution. Because of this symmetry, the Z-score formula is sometimes written as:

$Z=|x\text{-}\mu|\sigma$

Where the vertical lines in the equation means the absolute value of the number.

What the standardizing formula is really doing is computing the number of standard deviations X is from the mean of its own distribution. The standardizing formula and the concept of counting standard deviations from the mean is the secret of all that we will do in this statistics class. The reason this is true is that **all** of statistics boils down to variation, and the counting of standard deviations is a measure of variation.

This formula, in many disguises, will reappear over and over throughout this course.

The final exam scores in a statistics class were normally distributed with a mean of 63 and a standard deviation of five.

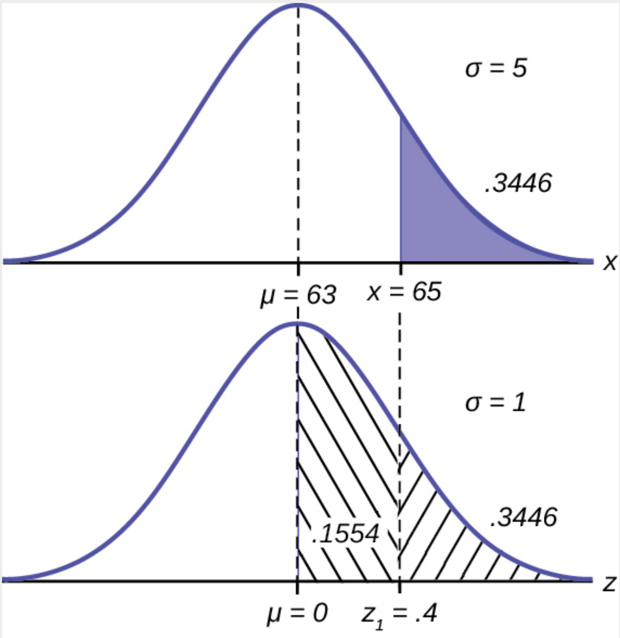
a. Find the probability that a randomly selected student scored more than 65 on the exam.
b. Find the probability that a randomly selected student scored less than 85.

a. Let X = a score on the final exam. $X \sim N(63, 5)$, where $\mu = 63$ and $\sigma = 5$

Draw a graph.

Then, find $P(x > 65)$.

$P(x > 65) = 0.3446$



$Z1 = x1 - \mu\ \sigma = 65-63\ 5 = 0.4$

$P(x \geq x_1) = P(Z \geq Z_1) = 0.3446$

The probability that any student selected at random scores more than 65 is 0.3446. Here is how we found this answer.

The normal table provides probabilities from zero to the value Z_1. For this problem the question can be written as: $P(X \geq 65) = P(Z \geq Z_1)$, which is the area in the tail. To find this area the formula would be 0.5 – $P(X \leq 65)$. One half of the probability is above the mean value because this is a symmetrical distribution. The graph shows how to find the area in the tail by subtracting that portion from the mean, zero, to the Z_1 value. The final answer is: $P(X \geq 63) = P(Z \geq 0.4) = 0.3446$

$z = 65 - 63\ 5 = 0.4$

Area to the left of Z_1 to the mean of zero is 0.1554

$P(x > 65) = P(z > 0.4) = 0.5 - 0.1554 = 0.3446$

b.

Z=x-μσ=85-635=4.4 which is larger than the maximum value on the Standard Normal Table. Therefore, the probability that one student scores less than 85 is approximately one or 100%.

A score of 85 is 4.4 standard deviations from the mean of 63 which is beyond the range of the standard normal table. Therefore, the probability that one student scores less than 85 is approximately one (or 100%).

Try It

The golf scores for a school team were normally distributed with a mean of 68 and a standard deviation of three.

Find the probability that a randomly selected golfer scored less than 65.

normalcdf(1099,65,68,3) $=$ 0.1587

A personal computer is used for office work at home, research, communication, personal finances, education, entertainment, social networking, and a myriad of other things. Suppose that the average number of hours a household personal computer is used for entertainment is two hours per day. Assume the times for entertainment are normally distributed and the standard deviation for the times is half an hour.

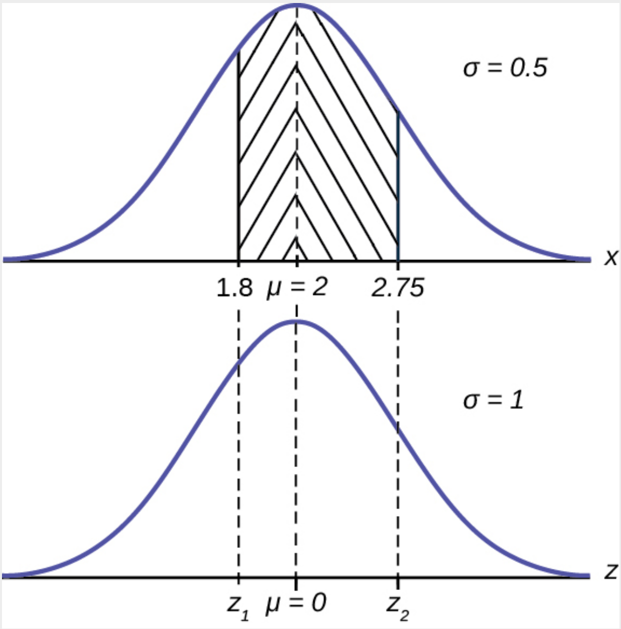
a. Find the probability that a household personal computer is used for entertainment between 1.8 and 2.75 hours per day.

a. Let $X =$ the amount of time (in hours) a household personal computer is used for entertainment. $X \sim N(2, 0.5)$ where $\mu = 2$ and

$\sigma = 0.5$.

Find $P(1.8 < x < 2.75)$.

The probability for which you are looking is the area **between** $x = 1.8$ and $x = 2.75$. $P(1.8 < x < 2.75) = 0.5886$

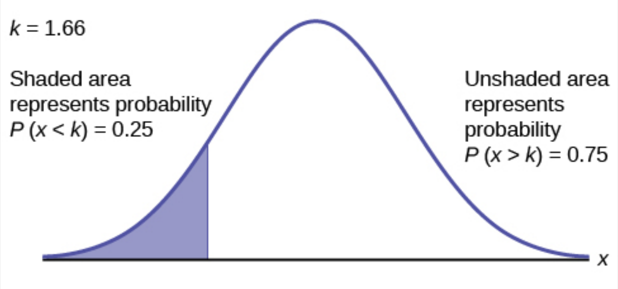


$P(1.8 \le x \le 2.75) = P(Z_i \le Z \le Z_2)$

The probability that a household personal computer is used between 1.8 and 2.75 hours per day for entertainment is 0.5886.

b. Find the maximum number of hours per day that the bottom quartile of households uses a personal computer for entertainment.

b. To find the maximum number of hours per day that the bottom quartile of households uses a personal computer for entertainment, **find the 25th percentile,** k, where $P(x < k) = 0.25$.



f(Z)=0.5-0.25=0.25, *therefore* $Z \approx$-0.675(*or just 0.67 using the table*)Z=x-μσ=x-20.5=-0.675, *therefore* x=-0.675*0.5+2=1.66 *hours.*

The maximum number of hours per day that the bottom quartile of households uses a personal computer for entertainment is 1.66 hours.

Try It

The golf scores for a school team were normally distributed with a mean of 68 and a standard deviation of three. Find the probability that a golfer scored between 66 and 70.

normalcdf(66,70,68,3) = 0.4950

There are approximately one billion smartphone users in the world today. In the United States the ages 13 to 55+ of smartphone users approximately follow a normal distribution with approximate mean and standard deviation of 36.9 years and 13.9 years, respectively.

a. Determine the probability that a random smartphone user in the age range 13 to 55+ is between 23 and 64.7 years old.

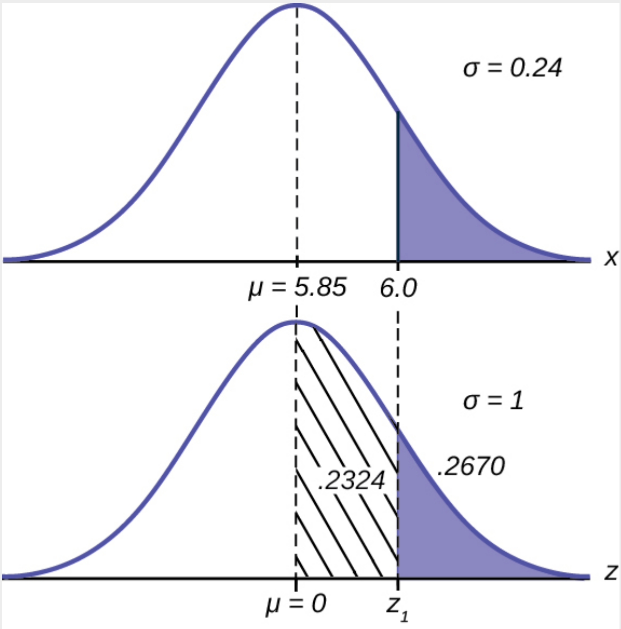
a. 0.8186

b. Determine the probability that a randomly selected smartphone user in the age range 13 to 55+ is at most 50.8 years old.

b. 0.8413

A citrus farmer who grows mandarin oranges finds that the diameters of mandarin oranges harvested on his farm follow a normal distribution with a mean diameter of 5.85 cm and a standard deviation of 0.24 cm.

a. Find the probability that a randomly selected mandarin orange from this farm has a diameter larger than 6.0 cm. Sketch the graph.



$Z1 = 6 - 5.85 \; .24 = .625$

$P(x \geq 6) = P(z \geq 0.625) = 0.2670$

b. The middle 20% of mandarin oranges from this farm have diameters between ____ and ____.

f (Z) = 0.20 2 = 0.10 , *therefore Z* ≈ ± 0.25 Z = x-μ σ = x - 5.85 0.24 = ± 0.25 → ± 0.25 * 0.24 + 5.85 = (5.79 ,5.91)

References

“Naegele’s rule.” Wikipedia. Available online at http://en.wikipedia.org/wiki/Naegele's_rule (accessed May 14, 2013).

“403: NUMMI.” Chicago Public Media & Ira Glass, 2013. Available online at http://www.thisamericanlife.org/radio-archives/episode/403/nummi (accessed May 14, 2013).

“Scratch-Off Lottery Ticket Playing Tips.” WinAtTheLottery.com, 2013. Available online at http://www.winatthelottery.com/public/department40.cfm (accessed May 14, 2013).

“Smart Phone Users, By The Numbers.” Visual.ly, 2013. Available online at http://visual.ly/smart-phone-users-numbers (accessed May 14, 2013).

“Facebook Statistics.” Statistics Brain. Available online at http://www.statisticbrain.com/facebook-statistics/(accessed May 14, 2013).

Practice questions

A local bank has determined that the daily balances of the chequing accounts of their customers are normally distributed with a mean of $280 and a standard deviation of $20.

1. What percentage of their customers has daily balances less than $290?
2. What percentage of their customers has daily balances between $250 and $275?
3. What percentage of their customers has daily balances over $260?
4. The Bank is planning a special promotion where it is rewarding its customers whose balances are in the top 15% with a free toaster. What account balance must a customer achieve in order to qualify for a free toaster?
5. 68.26% of balances will be between what amount?
6. What is the interquartile range for the account balances?

---

1. 0.6915
2. 0.3345
3. 0.8413
4. $300.70
5. $260 to $300
6. IQR = 293.5-266.5 = $27

The Old Baldy Tire Company is introducing a new steel belted radial tire. Old Baldy's engineering department has estimated that the average life of this tire will be 50,000 km and that the standard deviation of these tires will be 5000 km. It is assumed that the useful life of these tires follows a normal distribution.

1. What is the probability that:

    1. These tires will last for longer than 60,000 km?
    2. These tires will last for less than 38,000 km?
    3. These tires will last for between 45,000 and 58,000 km?
    4. These tires will last for between 39,000 and 43,000 km?

2. The Old Baldy Tire Company is considering offering a tire guarantee that each new set of tires will last a certain number of kilometers. If the tires fail to last the specified number of kilometers a new set of tires will be provided to the purchaser for free. The Old Baldy Tire Company wants to ensure that no more than 10% of the tires produced qualify for this guarantee. For how many kilometers should these tires be guaranteed to last?
3. 35% of tires will last less than how many kilometers?

1. 1. 0.0228
   2. 0.0082
   3. 0.7865
   4. 0.0669
2. 56407.8 km
3. 48073.4 km

Introduction to Sampling Distributions

Introduces the concept of sampling distributions.

When we take a random sample from a population, we expect that there is going to be some variability (i.e. sampling variability) between the information the sample gives us and the whole population. That is, we might find that the sample mean and the population mean are different. We may also find that if we take multiple random samples of size *n* that the sample mean for each sample is different. The following chapter looks at how we can better understand the sampling variability in statistics.

Before we go on, here is a reminder of a few terms and symbols.

A **parameter** is a descriptive measure of the population (eg. population mean, population standard deviation, population proportion).

A **statistic** is a descriptive measure of the sample (eg. sample mean, sample standard deviation, sample proportion).

Measure	Population	Sample

Sample size	*N*		*n*	
Mean	μx		x	
Standard deviation	σx		sx	
Proportion	π		p	>

Table of important symbols

The population mean, population standard deviation, and sample standard deviation have a subscript of *x* to demonstrate that they are the measure for the variable *X*. Though this is mostly notational, it does become important later in this chapter.

Number of women in each sample of size 100 This is based off of calculations on how long it would take a network of supercomputer in 2011 to work through all possible combinations of a 256 bit encryption. By the way, there are less possibilities in a 256 bit encryption than there are all possible samples of size 100 from a population of 12,000.

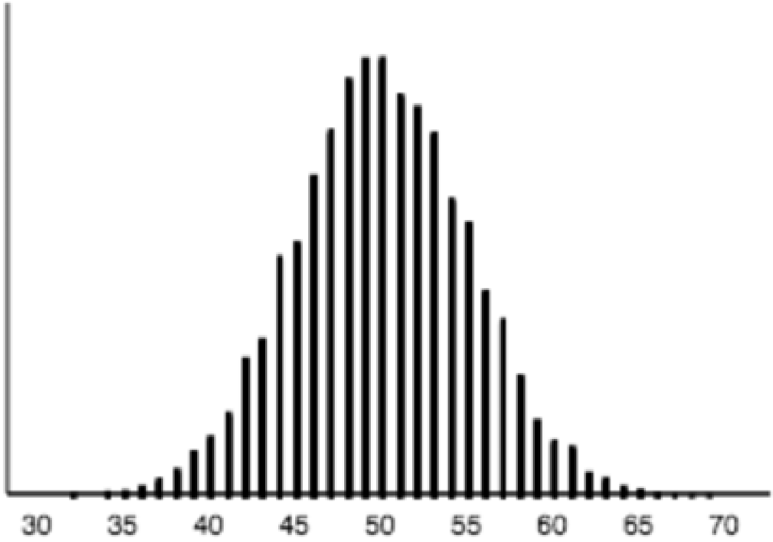
What is a sampling distribution?

Suppose we take many different random samples of 100 university students from a university that has an equal number of men and women.

The number of women will vary amongst the samples. For example, one sample could have 45 women, another sample could have 48 women, another sample could have 52 women, etc.

Though it could be possible that we get a random sample that only has 2 women in it, it would be pretty unlikely. Instead, we would expect that most of the samples would have around 50 women in it with some variation around that value.

Figure 1 is the result of a simulation that took 10,000 samples of size 100 from a population that had an equal about of women and men. The horizontal axis is the number of women in each sample. The height of each bar is the number of samples that had that many women.



Notice how the most common number of women is around 50 (i.e. the average), but there is variation from that 50. Most samples have between 40 and 60 women.

The variability among random samples of size *n* from the same population is called **sampling variability**.

A probability distribution that characterizes some aspect of sampling variability is termed a **sampling distribution**. A sampling distribution is constructed by taking all possible samples of a size *n* from a population. Then for each sample, a statistic is calculated (e.g. sample mean, sample proportion, sample standard deviation). The sampling distribution is then created by making a graph of all of these samples.

Actually constructing a sampling distribution is often very difficult. A medium sized university in Canada might have 12,000 students. All possible samples of size 100 from that population would result in 5.87×10249 unique samples! Think about that. One billion is 109 . Google is named after a googol (10100) because they wanted Google to be associated with an immense amount of data. Yet a googol is smaller than all possible samples at 100 from the medium sized university. If we got a computer to find all possible samples, it would take it over a billion years to find them[footnote]! Therefore, actually constructing a true sampling distribution in most situations is incredibly hard, incredibly time consuming, and not really worth it. Thus when we talk about sampling distributions, we talk about a **theoretical sampling distribution**.

That is, we theorize what this sampling distribution would look like if it was possible to examine all possible samples.

Due to these limitations, we often look at an empirical sampling distribution, instead of a theoretical sampling distribution. An **empirical sampling distribution** is created by taking many samples from a population and finding a statistic for each sample, but not doing this for all possible samples. The plot shown in is an example of an empirical sampling distribution as it only contains 10,000 samples and not all possible samples. The statistic in is the number of women, but we could have also looked at the proportion of women.

In summary, a sampling distribution is a distribution of a statistic. This differs from other distributions, like the population distribution, which are distributions for individual data values.

Why do we care about sampling distributions?

Suppose we take a random sample of 100 students from a medium sized university and we find that 75 of them are women. Does this call into question the assumption that 50% of the students are women? This is hard to figure out unless we know how likely it is that we could have found this random sample, assuming that there are an equal number of men and women.

The sampling distribution helps us find this probability. From the empirical sampling distribution in Figure 1 we can find the probability of getting a random sample of 75 women, assuming that there are an equal number of men and women is 0.0000%. That is, it is really unlikely to get a random sample of 75 women out of 100 if there are an equal number of men and women in the population. Based on this, we can be fairly confident that this university probably doesn't have an equal number of men and women. Instead, it is more likely that there are women than men at this university.

The process described above is called **inferential statistics**. Inferential statistics is used to make a conclusion about the population (all students at the university) from a sample (100 students). In general, to do any form of inferential statistics, we need to use a sampling distribution to either determine how likely or unlikely a statistic is (in hypothesis testing) or to estimate a parameter from a statistic (confidence intervals).

Thus sampling distributions are the backbone of inferential statistics.

Note: What was described above about the proportion of women at a university should sound familiar. In Chapter 4, we used the binomial distribution to determine how not unlikely or unlikely events were. The binomial distribution was helping us understand the sampling distribution of proportions.

Constructing Empirical Sampling Distributions
Introduction to how to construct a sampling distribution.

How to construct an emprical sampling distribution

If we have access to the population, we can construct an empirical distribution from it. This can be done by using computer software to pull random samples from a population. An example of one such tool is from the Rossman Chance website, which has an applet that allows you to create an empirical sampling distribution from a finite population: http://www.rossmanchance.com/applets/OneSample53.html

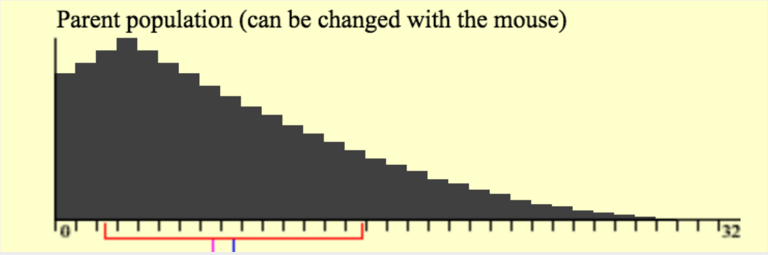
When constructing an empirical sampling distribution, it is important to keep the law of large numbers in mind. That is, the more samples you take, the closer the empirical sampling distribution will be to the theoretical sampling distribution. In general, empirical sampling distributions should be constructed from at least 10,000 samples.

To get an idea of how an empirical sampling distribution is constructed, go to http://onlinestatbook.com/stat_sim/sampling_dist/index.html

The images/figures in this example were generated from David Lane's sampling distribution applet that is part of the OnlineStatBook project [footnote]. Online Statistics Education: A Multimedia Course of Study (http://onlinestatbook.com/). Project Leader: David M. Lane, Rice University.

Figure 1 shows the histogram of the population we are going to generate an empirical sampling distribution from. We call this population the **parent population** as it is the population we are creating the sampling distribution from. Notice that the parent population is skewed left.

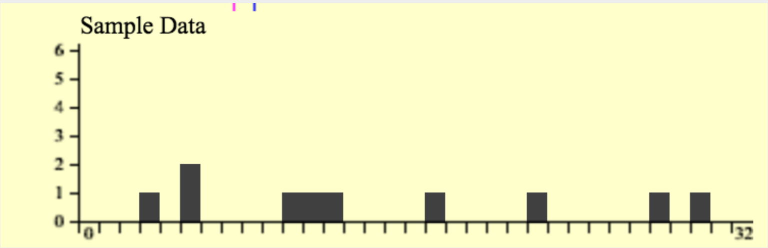
Parent population



We are going to take multiple samples of size 10 from the parent population and look at the statistic of the sample mean for each sample.

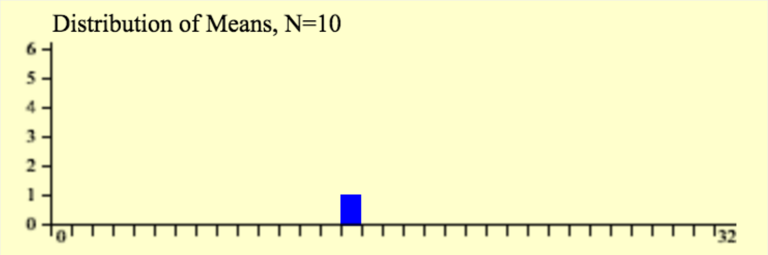
Here is the first sample:

Sample of size 10 from the parent population



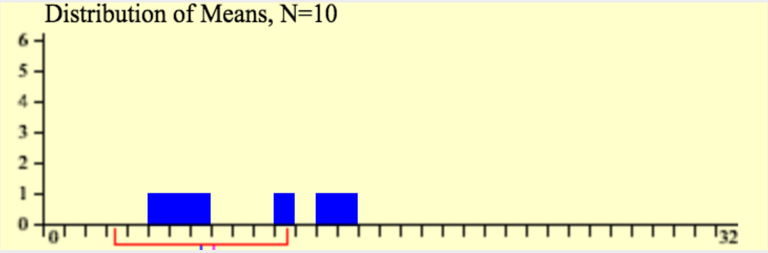
This is the sample mean of the sample:

Sample mean for one sample of size 10



Now one sample mean is not enough to tell us what the sampling distribution looks like. So let's take a few more samples. Let's take 5 more samples of size 10 and plot their sample means:

Six sample means from parent population

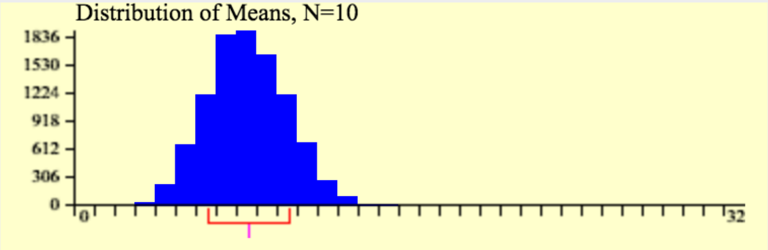


This is still a pretty small sample.

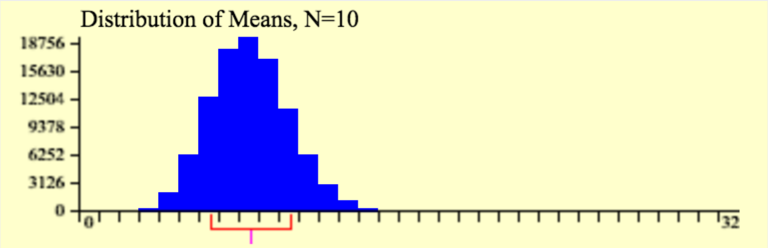
There are two sample sizes here. One is the size of the sample we are taking from the parent population (10). The other is the number of samples we've taken (6). The first is the sample size for the sample. The second is the sample size for the empirical sampling distribution.

Now let's take 10,000 samples of size 10 from the population and plot each of their sample means.

This is what we get:
10,000 sample means from parent population



Finally, let take 100,000 samples of size 10 from the population and plot each of their sample means. This is what we get:
100,000 sample means from parent population



Notice how there is no real difference between the distributions (shape, centre and variation) in Figure 5 and Figure 6. This means that are empirical distribution is now giving us a good sense of what the theoretical sampling distribution would look like. When this happens, this is called **convergence**. That is, the empirical sampling distribution is converging on the theoretical sampling distribution. As the sample size of the empirical sampling distribution increases this is expected to happen due to law of large numbers.

Bootstrapping

Suppose we don't have access to the population. This can happen if the population is infinite (e.g. in a manufacturing process) or where the population is large (e.g. population of Canada) or where most researchers wouldn't have access to the population (e.g. list of students at a university). Can we still construct an empirical sampling distribution?

The answer is yes! To do this, we use a process called **bootstrapping**. Essentially bootstrapping follows the same procedure as outlined in Example 1, but instead of using a parent population, we use a parent sample. That is, we take a good sample from the population and use that to construct the sampling distribution.

Again the law of large numbers applies. If the random sample from the population is large enough, then the sample will most likely be a good estimate of the population. Then the empirical sampling distribution generated by the sample will most likely be a good estimate of the theoretical sampling distribution of the population.

Bootstrapping only works if the sample being used has been collected properly and that the sampling technique ensures that the sample is random, the sample is representative of the population, and the sample size is large enough. There are no set rules on how big the sample needs to be, but for bootstrapping the bigger the better.

The Central Limit Theorem
Introduction to the key properties of the central limit theorem for the mean and proportions.

Another way to determine what the sampling distribution looks like is by using theory. The main theory that helps us understand the characteristics of the sampling distribution is called the **central limit theorem**.

The central limit theorem is an incredibly useful and powerful theorem. The theorem tells us about the distribution of many different sampling distributions. But be careful! The central limit theorem cannot be applied always and only applies to sampling distributions.
 This formula assumes the population is infinite or very large. If this is not the case, then the formula is
 As the population size (*N*) increases, N − nN − 1 approaches 1 and no longer affects the standard error.
 The images/figures that follow were generated from David Lane's sampling distribution applet that is part of the OnlineStatBook project
Online Statistics Education: A Multimedia Course of Study (http://onlinestatbook.com/). Project Leader: David M. Lane, Rice University.
Parent populationSampling distribution for Figure 1 for samples of size 2Sampling distribution for Figure 1 for samples of size 5 Sampling distribution for Figure 1 for samples of size 10 Sampling

distribution for Figure 1 for samples of size 16 Sampling distribution for Figure 1 for samples of size 20 Sampling distribution for Figure 1 for samples of size 25

The central limit theorem for the sampling distribution for sample means

The sampling distribution for the sample means comes from a parent population that is comprised of quantitative data. Random samples of size *n* are taken from the parent population and the sample mean is calculated for each sample. What will the distribution of the sample means look like? That is, what is the shape of the distribution of sample means, where are the sample means centred, and what is the sampling variability?

The following refers to the theoretical sampling distribution for the sample means. Further, when sample size is mentioned, it is referring to the size of the sample taken from the population. That is, it is not referring to how many different random samples have been taken.

Where are the sample means centred?

As the sample means are estimating the population mean, it makes sense that the sample means are centred around the population mean.

In the previous section, we saw the right skewed parent population in Figure 1. The population mean of that parent population is 8.08. Notice that the empirical sampling distributions shown in Figures 5 and 6 are both centred around 8.08.

In general, the mean of the theoretical sampling distribution for the sample means equals the population mean.

$$\mu_{\bar{x}} = \mu_x$$

The variable for the sample means is $\bar{x}$. That is why the subscript for the mean of the sample means ($\mu_{\bar{x}}$) has changed.

What is the sampling variability? (or what is the variation in the sampling distribution)

Based on the law of large numbers, the sampling variability of the sample means will decrease as the sample size increases. As the sample size increases, the sample means will become better and better estimates of the population mean and, therefore, there will be less variability between them. That is, there will be more variability between the sample means for samples of size 2, then there will be for samples of size 30.

Just like we can measure variability for individual data values, we can also measure variability for sample means. We will use the standard deviation to measure the sampling variability. The standard deviation of the sampling distribution for sample means is called the **standard error of the sample means**. It is found with the following formula [footnote] :

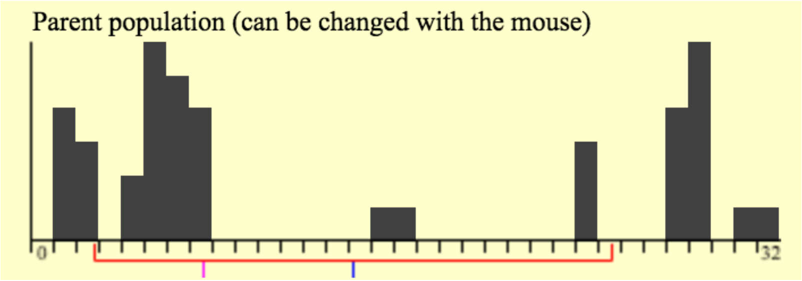
$$\sigma \bar{x} = \sigma n N - n N - 1$$

$$\sigma \bar{x} = \sigma n$$

What is the shape of the distribution?

This is actually a really interesting question.

Suppose the parent population looks like this [footnote]:

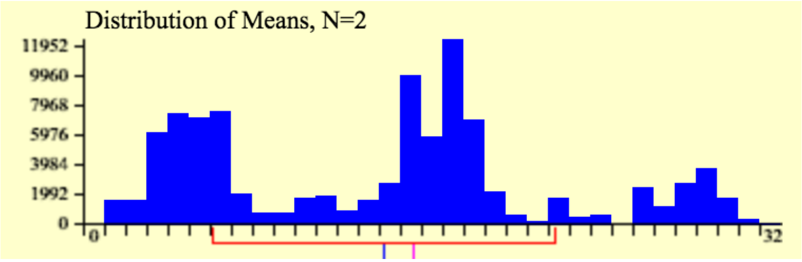


What will the sampling distribution for sample means look like?

Here’s the answer:

- If the parent population is normal, then the sampling distribution for sample means will be normal. Always.
- As the sample size of the samples being taken from the parent population increases, the more normal the sampling distribution for sample means will become.

Since the population in Figure 1 is not normally distributed, then we would expect the sampling distribution will not be normal for smaller sample sizes, but will be normal for larger sample size.



For each of these empirical sampling distributions, 100,000 samples were taken of size **n**. Therefore, we can be very confident that the empirical sampling distributions are good representations of the theoretical sampling distributions.

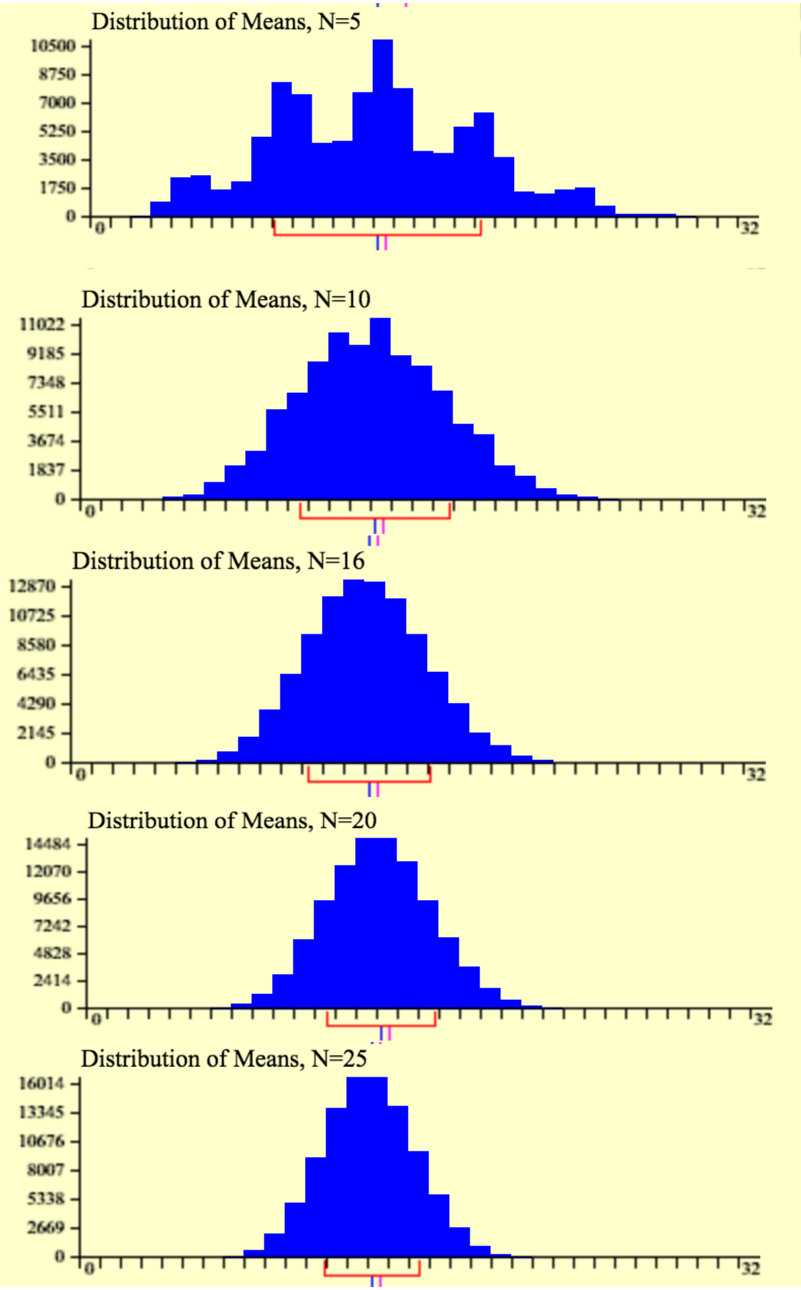


Figure 1 (the parent population) is not even close to being normal, but notice that as the sample size increases, the sampling distribution for sample means gets closer and closer to being normally distributed!

In general, the closer the population is to being normally distributed, the “faster” the sampling distribution gets closer to normal. Here “faster” means for a smaller sample size.

The central limit theorem states that regardless of the shape of the population, if the sample size is greater than 30, the sampling distribution will be approximately normal.

Measure	Population	Sample	Sampling distribution for the sample mean
Mean	μx	$\bar{x}$	$\mu \bar{x} = \mu x$
Standard deviation	σx	sx	$\sigma \bar{x} = \sigma n$ (standard error)

The central limit theorem for the sampling distribution for sample proportions

The sampling distribution for the sample proportions comes from a parent population that satisfies the criteria of the binomial distribution. Random samples of size *n* are taken from the parent population and the sample proportion is calculated for each sample. What will the distribution of the sample means look like? That is, what is the shape of the distribution of sample proportions, where are the sample proportions centred, and what is the sampling variability?

The sampling distribution for sample proportions has similar characteristics as the sampling distribution for the sample means.

Where are the sample proportions centred?

They are centred around the population proportion.

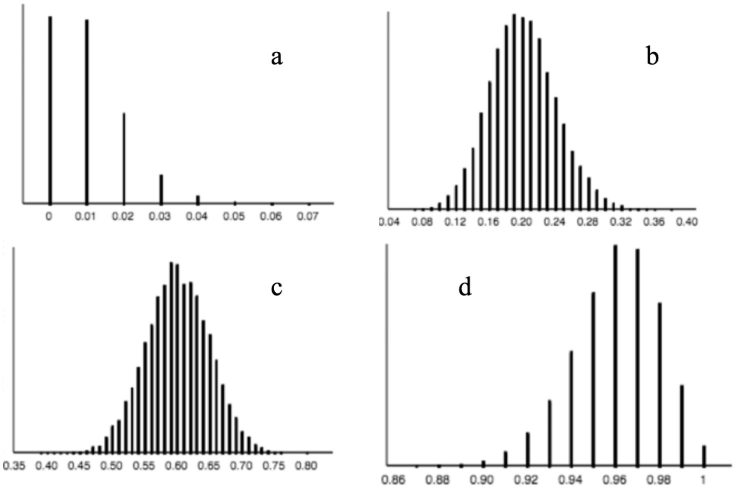
What is the sampling variability?

It decreases as the sample size increases.

What is the shape?

The shape of sampling distributions of the sample proportions also becomes normal. Unlike for sample means though, the normality is not based on sample size, but is based on the number of successes ($n\pi$) and failures ($n(1-\pi)$).

To illustrate, here are the empirical sampling distributions for proportions for various population proportions. The sample size is 100 in each case and the number of samples taken is 10,000.



In Figure 8 a, $n = 100$ and $\pi = 0.01$. Therefore, the number of successes is 1 and the number of failures is 99. The sampling distribution is skewed to the right.

In Figure 8 b, $n = 100$ and $\pi = 0.20$. Therefore, the number of successes is 20 and the number of failures is 80. The sampling distribution is approximately normal.

In Figure 8 c, $n = 100$ and $\pi = 0.60$. Therefore, the number of successes is 60 and the number of failures is 40. The sampling distribution is approximately normal.

In Figure 8 d, $n = 100$ and $\pi = 0.96$. Therefore, the number of successes is 96 and the number of failures is 4. The sampling distribution is skewed to the left.

In general, the shape of the sampling distribution for sample proportions is approximately normal if the number of successes and the number failures are both at least 5.

If the sampling distribution for sample proportions is normal, we can find probabilities for the distribution using two methods. The first method is using the binomial distribution. The second method is the normal distribution. This might seem a bit strange as the binomial distribution is for discrete random variables and the normal distribution is for continuous random variables. In reality, we use the normal distribution to approximate probabilities for the sampling distribution for sample proportions. This is called the **normal approximation to the binomial distribution**. To get the exact probability, one would need to use the binomial distribution. But this can be cumbersome when the sample sizes are very large (e.g. 1000). Therefore, using the normal distribution can be beneficial, especially because it gives very accurate approximations. In example 6.4 below we will investigate this further.

Further when we begin to do inferential statistics, we won't know the population proportion (otherwise inferential statistics wouldn't be necessary). Since we won't know π it will hard to use the binomial distribution. Therefore, we use the normal approximation to the binomial distribution instead.

If we use a normal approximation to the binomial distribution, we need to know the mean and standard deviation of the sampling distribution.

The mean of the sampling distribution for sample proportions is the population proportion.
$\mu_{\hat{p}} = \pi$

The standard deviation of the sampling distribution for sample proportions (or the **standard error of sample proportions**) is found using the following formula:

$$\sigma\ p\ \hat{}\ = \pi(1-\pi)n$$

Calculating Probabilities for Sampling Distributions--A Series of Examples

Four examples that show how to calculate and interpret probabilities found for sampling distributions

If a sampling distribution is normally distributed, then we can find probabilities for the sampling distribution using the normal distribution just like we did in Chapter 5.

The z-score for the sampling distribution for sample means would be:

$Z = \bar{X} - \mu \bar{x}\ \sigma \bar{x} = \bar{X} - \mu x \sigma x n$

The z-score for the sampling distribution for sample proportions would be:

$Z = \hat{p} - \mu \hat{p}\ \sigma \hat{p} = \hat{p} - \pi \pi (1 - \pi) n$

The Old Baldy Tire Company is introducing a new steel belted radial tire. Old Baldy's engineering department has estimated that the mean life of this tire will be 50,000 km and that the standard deviation of these tires will be 10,000 km. Suppose a large number of random samples of 100 tires is taken. The shape of the population distribution is unknown.

1. Can we assume the distribution of the mean life of these tires will be normal?

Explain.

2. Regardless of your result in a), assume that we are dealing with a normal distribution. Find the probability that the mean life of a random sample of 100 tires is less than 49,000km.
3. A competitor of Old Baldy's takes a random sample of 100 tires and finds their mean life to be 49,000 km. Based off of this data, they claim that the engineering department of Old Baldy's has exaggerated the mean life of their new tires. Do you support the competitor's claim? Explain.

---

1. Yes. As the sample size is greater than 30 (it is 100), we can assume that the sampling distribution of the sample mean lifespan of the tires is normally distributed regardless of the shape of the sampling distribution due to the central limit theorem.
2. $\bar{x}$ = mean lifespan of 100 Old Baldy tires, $\mu \bar{x} = \mu x = 50,000$, $\sigma \bar{x} = \sigma n = 10,000/10 = 1000$. Since we know that the data is normally distributed, we can use a computer program to calculate the probability $P(\bar{X} < 49,000)$. From the computer program, we get $P(\bar{X} < 49,000) = 15.87\%$
3. No. The probability that a random sample

> of 100 Old Baldy tires has a mean lifespan of 49,000km is 15.87% (assuming Old Baldy's claim). This means that this event is likely to occur (as it is greater than 10%), under the assumption that the tires last on average 50,000 km, and does not provide sufficient evidence against Old Baldy's claim.

The maintenance manager at a popular mountain resort is trying to determine if the aging gondola is in need of some renovation—or perhaps outright replacement. Right now, the maximum load of the gondola is 900 kilograms or 12 persons. The manager knows that the average weight of North Americans has been on the rise for several years and wishes to test what the probabilities might be of this gondola exceeding the maximum capacity.

Since the operators don't currently look at gender—just numbers—the manager is concerned about what might happen if the worst-case scenario were to occur: 12 large adult males were allowed on the gondola at the same time.

To investigate this further the manager did some research into the current average weight of adult males and discovered that it is about

80 kilograms. He also knows that adult weight tends to be normally distributed by gender, with a standard deviation for males of about 12 kilograms.

1. Given this information, he first wants to know what the individual weight allowance is (i.e. the per person average) that the gondola can withstand.
2. He also wants to know how likely is it that the individual weight of any randomly selected male will exceed the individual weight allowance calculated above.
3. Finally, he wants to know how likely it would be that the average weight of a sample of 12 adult males would exceed the average individual weight allowance.
4. Based on your answers, do you think the manager should renovate the gondola? Is there any further information that the manager would need?

---

1. 900kg/12 people = 75kg/person
2. Since this is about an individual, we will use µx and σx . As stated in the question, we know that the population is normally distributed. From this, a computer program calculated that $P(X \geq 75) = 66.15\%$, where X is the weight of an individual person on the gondola.

3. Now, we are being asked about the mean for 12 people. Therefore, this question is about finding a probability for the sampling distribution for sample means. Therefore, we will use $\bar{x}$ = mean weight of 12 people, $\mu \bar{x} = \mu x = 80$, $\sigma \bar{x} = \sigma n = 12/3.46 = 3.46$. Since the population distribution is normal, we know that the sampling distribution will also be normal (regardless of the sample size). Therefore, we can use a computer program to calculate the probability $P(\bar{X} > 75)$. We get 92.55%.
4. The probability found in c) is the probability that the average mass of 12 adult males will exceed the maximum individual weight for the gondola. The next question is “how likely is it that there will be 12 adult males on the gondola?” The manager should do further research to determine this before making a decision. While waiting for the results, the manager should implement a policy where any large groups of males are broken up and are required to take the lift in separate groups. I.e. break up a group of 12 males into two groups of 6 males.

The city of Montreal has an extensive bike lane system. In fact, it is one of the largest in North

America. But many cyclists find that even with all of the bike lanes, it is still hard to get around the city on a bike. In particular, there are many lanes that run east/west, but few that run north/south. Thus, they are encouraging the city council to focus less on adding lots of kilometers to the system, but instead making sure that the current system properly connects all parts of the city.

The city council will only go forward with this idea if at least 66% of the residents support focusing on connecting the system rather than expanding the system.

Suppose that 62% of residents do support connecting the system rather than expanding it. What is the probability that a random sample of 1000 residents will have a sample proportion of at least 66%?

1. Find the above probability using the binomial distribution.
2. Find the above probability by using the sampling distribution for sample proportions.
3. Compare the two answers. Do they give similar answers?
4. Based on your answers, do you think that it is possible that the city of Montreal will choose to focus on connecting the bike

path system?

---

1. Since we are using the binomial distribution, we are being asked to find the probability that at least 660 of the 1000 people in the poll will want to focus on connecting the system. The 660 comes from 66% of 1000. In other words, we are asked to find the $P(X \geq 660)$, with n = 1000 and π = 60%. Using a computer program, this yields a probability of 0.48%. This is found, by highlighting all of the values above 660 and including 660.
2. Since we are using the sampling distribution for sample proportions, we are asked to find the probability that the sample proportion will be at least 66%. In other words, we are asked to find the $P(\hat{p} \geq 0.66)$. We can assume the sampling distribution for sample proportions is normal as the number of successes ($n\pi = 1000 \times 0.62 = 620$) and the number of failures ($n(1-\pi) = 1000 \times 0.38 = 380$) are both at least five. Therefore, we will use the normal distribution to find the probability with $\mu_{\hat{p}} = \pi = 0.62$ and $\sigma_{\hat{p}} = \pi(1-\pi)n = 0.62(1-0.62)1000 =$ 0.01475. Therefore, using a computer program we find $P(\hat{p} \geq 0.66) = 0.33\%$
3. The two probabilities are quite close. They

are only 0.15% apart. Therefore, the two methods give us similar results.

4. It is unlikely that if the proportion of residents that want to focus on connecting the bike system is 62% that a poll of 1000 people would result in a sample proportion of 66%. Therefore, it is unlikely that the city of Montreal will chose to focus on connecting the system.

Video games are gaining more and more popularity. Children often try to convince their parents to buy games even when they are not appropriate. For example, they may want to play a very violent game that is not appropriate for their age group. To help parents out, video games have rating categories to suggest age appropriateness. But how aware are parents of these categories?

To investigate this, you conduct a survey of Canadian families that have young children who play video games. You show parents three video game covers that have the category rating clearly marked on it. You then ask the parents whether the games would be appropriate for children and why. If the parent correctly identifies which games are appropriate for their children and refers to the ratings in making their choice, you categorize the parents as well informed.

Suppose that we want to use your results to justify the claim that less than 30 percent of parents are well informed about video game ratings. In your random sample of 1000 parents, you actually found that 27 percent of the parents that you polled were well informed about video game ratings.

1. Assuming that the proportion of parents that are well informed about video game ratings is 30%, what is the probability that you would observe a sample proportion of less than 27%. Use the normal approximation of the binomial distribution to find your answer.
2. Based on your results, do you believe that this is enough evidence to suggest that less than 30% of parents are well informed about video game ratings? Explain your answer.

---

1. Since we are using the sampling distribution for sample proportions, we are asked to find the probability that the sample proportion will be at most 27%. In other words, we are asked to find the *P*($\hat{p} \leq 0.27$). We can assume the sampling distribution for sample proportions is normal as the number of successes (and

the number of failures (($n\pi = 1000 \times 0.30 = 300$) and the number of failures ($n(1-\pi) = 1000 \times 0.7 = 700$) are both at least five. Therefore, we will use the normal distribution to find the probability with $\mu \hat{p} = \pi = 0.30$ and $\sigma \hat{p} = \pi(1-\pi)n = 0.3(1-0.3)1000 = 0.01145$. Therefore, using a computer program we find $P(\hat{p} \leq 0.27) = 0.44\%$.

2. Since the probability that we would observe a sample proportion of 27% (assuming a population proportion of 30%) is 0.44%, it is very unlikely we would have observed this evidence if the assumption is true. Therefore, it is more likely that the population proportion is less than 30%. Thus there is enough evidence to suggest that less than 30% of parents are well informed about video game ratings.

Practice questions

The following practice questions are from Lyryx Learning, Business Statistics I -- MGMT 2262 -- Mt Royal University -- Version 2016 Revision A. OpenStax CNX. Sep 8, 2016 http://cnx.org/contents/f3aefa9e-58d2-41ea-969f-04dc2cb04c82@5.5

Use the following information to answer the next ten exercises: A manufacturer produces 25-pound lifting weights. The lowest actual weight is 24 pounds, and the highest is 26 pounds. Each weight is equally likely so the distribution of weights is uniform. A sample of 100 weights is taken. The standard deviation is 0.58 pounds.

1. What is the distribution for the weights of one 25-pound lifting weight? What is the mean and standard deivation?
2. What is the distribution for the mean weight of 100 25-pound lifting weights?
3. Find the probability that the mean actual weight for the 100 weights is less than 24.9.

---

1. Uniform with a mean of 25 and a standard deviation of 0.58 pounds. Remember when a distribution is uniform all of the values are equally likely. Therefore the mean will be halfway between the lowest value (24) and the highest value (26).
2. Normal with a mean of 25 and a standard deviation of 0.0577
3. 0.0416

Find the probability that the mean actual weight for the 100 weights is greater than 25.2.

0.0003

Find the 90th percentile for the mean weight for the 100 weights.

25.07

Suppose that the distance of fly balls hit to the outfield (in baseball) is normally distributed with a mean of 250 feet and a standard deviation of 50 feet. We randomly sample 49 fly balls.

1. What is the probability that the 49 balls traveled an average of less than 240 feet?
2. Find the 80th percentile of the distribution of the average of 49 fly balls.

---

1. 0.0808
2. 256.01 feet

According to the Internal Revenue Service, the average length of time for an individual to complete (keep records for, learn, prepare, copy, assemble, and send) IRS Form 1040 is 10.53 hours (without any attached schedules). The distribution is unknown. Let us assume that the standard deviation is two hours. Suppose we randomly sample 36 taxpayers.

1. Would you be surprised if the 36 taxpayers finished their Form 1040s in an average of more than 12 hours? Explain why or why not in complete sentences.
2. Would you be surprised if one taxpayer finished his or her Form 1040 in more than 12 hours? In a complete sentence, explain why.

---

1. Yes. I would be surprised, because the probability is almost 0.
2. No. I would not be totally surprised because the probability is 0.2312

Suppose that a category of world-class runners are known to run a marathon (26 miles) in an average of 145 minutes with a standard deviation of 14 minutes. Consider 49 of the races. Let X – the average of the 49 races.

1. Find the probability that the runner will average between 142 and 146 minutes in

these 49 marathons.

2. Find the 80th percentile for the average of these 49 marathons.
3. Find the median of the average running times.

---

1. 0.6247
2. 146.68
3. 145 minutes

Determine which of the following are true and which are false. Then, in complete sentences, justify your answers.

1. When the sample size is large, the mean of X – is approximately equal to the mean of X.
2. When the sample size is large, X – is approximately normally distributed.
3. When the sample size is large, the standard deviation of X – is approximately the same as the standard deviation of X.

---

1. True. The mean of a sampling distribution of the means is approximately the mean of the data distribution.
2. True. According to the Central Limit Theorem, the larger the sample, the closer

the sampling distribution of the means becomes normal.

3. The standard deviation of the sampling distribution of the means will decrease making it approximately the same as the standard deviation of X as the sample size increases.

Introduction to Confidence Intervals

Introduction to collection on confidence intervals

From Chapter 6, we know that if we take many samples of the same size from a population and calculate the sample means, the sample means will be clustered around the population mean, but many of them won't be exactly the same as the population mean. Therefore, we can estimate the population mean using a sample mean, but we expect there to be a certain amount of error in that estimate. To determine that error, we can look at the standard error. That is, we can look at the amount of variation between the sample means.

In the chapter, we will use this information about how sample means behave to help us make estimates about the population mean of unknown populations. We will also do this with sample proportions and population proportions. That is, the goal of this chapter is to make inferences about the population from sample data. This is our first foray into **inferential statistics**.

By the end of this section, the student should be able to

- Find and interpret confidence intervals that estimate the population mean and the population proportion.
- Understand the properties of the Student-t

distribution.

- For confidence intervals for the population mean, can determine whether to use the Student-t distribution or the standard normal distribution as a model.
- Find the minimum sample size needed to estimate a parameter given a margin of error.

What are Confidence Intervals?

Explanation of what confidence intervals are.

Suppose you are trying to determine the mean rent of a two-bedroom apartment in your town. You might look in the classified section of the newspaper, write down several rents listed, and average them together. This provides a point estimate of the true mean. If you are trying to determine the percentage of times you make a basket when shooting a basketball, you might count the number of shots you make and divide that by the number of shots you attempted. In this case, you would have obtained a point estimate for the true proportion.

A **point estimate** is a single value used to estimate a population parameter. For example, the sample mean is a point estimate of the population mean. But point estimates do not give a sense of how much error there is in an estimate. Thus, we instead want to provide an **interval estimate** for the population parameter takes into account error. The type of interval estimate we will learn about in this chapter is called a **confidence interval**.

From our work on sampling distributions, we know that the sample mean probably won't be exactly the population mean. Instead we expect it to be slightly larger or smaller than the population mean. But by how much? The **margin of error**, denoted E, measures how much we expect the statistic to vary from the parameter. The margin of error is computed by looking at how much variation is in the sampling distribution and the level of confidence (discussed below).

To calculate a confidence interval, you take the statistic and you add and subtract the margin of error from it. For example, if you are trying to estimate the population mean, you would take the sample mean and add and subtract the margin of error from it: $\bar{x}-E,\bar{x}+E$. This gives an interval of values that you expect the population mean to fall between.

A recent opinion poll asked Canadians their opinion of the work of the current Prime Minister of Canada. 53% of Canadians approved of his work with a margin of error of 2.6%. The statistic is a sample proportion of 53% and we are trying to estimate the true proportion of Canadians who approved of the Prime Minister's work. We know that there will be error in that estimate and it has been measured to be 2.6%. Therefore, we are estimating that the true proportion of all Canadians who approve of the Prime Minister's work is between 53% ± 2.6% or between 50.4% and 55.6%.

Though confidence intervals change depending on the sample, but the parameter being estimated is fixed. For example, on a specific day, the population mean rent of a two-bedroom apartment in your town is a specific value. You are trying to estimate it, but it is fixed. The confidence interval, on the other hand, changes depending on the sample you take. Suppose instead of looking at the classified section of a newspaper, you looked at a rental website. Then the sample might be different, which will result in a different confidence interval. Or suppose you stood outside a mall entrance and asked every fifth person what they paid in rent for their two-bedroom apartment, then your sample would be different, which will result in a different confidence interval. These three different confidence intervals are all estimating the same thing, the population mean rent of a two-bedroom apartment in your town, but since each of the samples are different, the sample means will be different which will result in different estimates. In short, the parameter being estimated is *not* a random variable. But the confidence interval being used to estimate the parameter varies depending on the random sample taken.

In the following sections, we will learn how to calculate the margin of error for the mean and proportion. For each situation, we will use a different model to find the margin of error. It should be noted that all of the models are based on the assumption that a random sample has been calculated. Therefore, finding a confidence interval based on the convenience sample of the rent in today's classified ads is not appropriate. This is important to remember when you are critically assessing a confidence interval provided to you. No matter how prettily the confidence interval is presented, if it was constructed from a non-random sample, it is useless. It is like baking an apple pie from rotten apples. It might look good, but it is still rotten.

100 confidence intervals generated from 100 random sample of the rent of two-bedroom apartments in your townOnline Statistics Education: A Multimedia Course of Study (http://onlinestatbook.com/). Project Leader: David M. Lane, Rice University.

Why is it called a confidence interval?

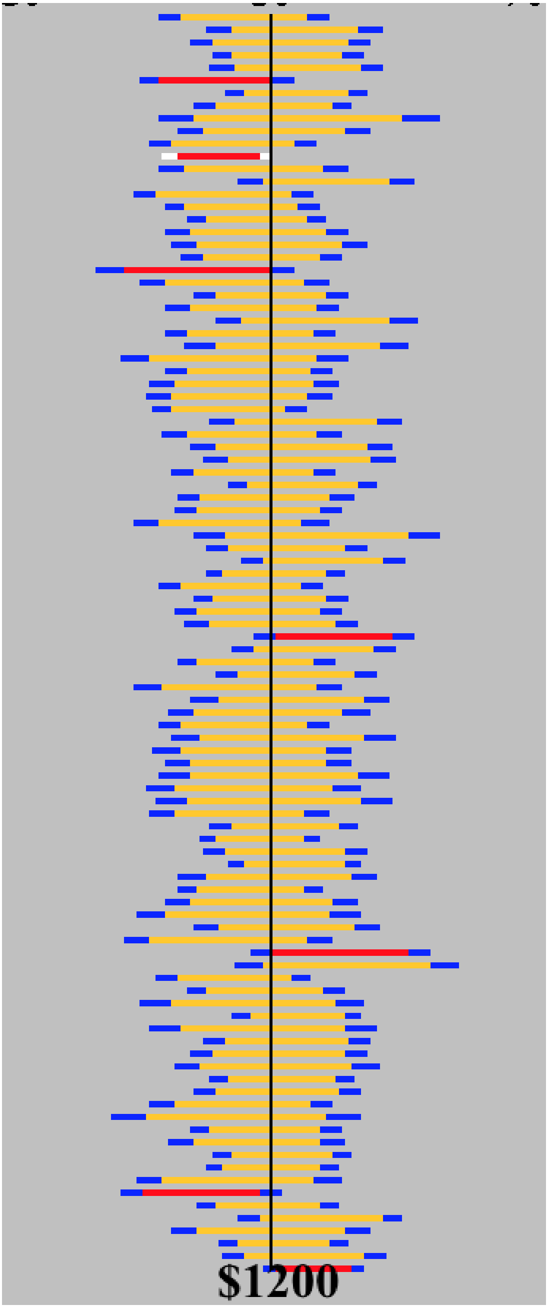
If you are trying to estimate how much it will cost to go on a trip to Montreal for five days, you can work out with strong confidence the cost of the flight and hotels, but then you have to start making estimates about how much food and entertainment will cost while you're there. You can get a pretty good estimate of what it will cost, but your friend who you are trying to convince to come with you might want to know how confident you are in that estimate. Are you the kind of person who just guesses at the cost of meals or did you look at restaurantsÕ menus to come up with a sense of what meals cost in Montreal? Did you take into account snacks? The cost of renting a car or taking the bus? Did you assume you were going to do an equal number of free and paid admission activities? All of this affects the confidence you have in your estimate.

For a confidence interval, it is much easier to determine how much confidence we have in our estimate because confidence intervals come with a **level of confidence** (or **confidence level**).

To understand the confidence level, let's go back to the two-bedroom apartment situation. Let's now suppose that 100 people on the same day were very curious about determining the mean rent for two-bedroom apartments in your town. Each of these 100 people went out and found their own random sample of fifty people who rent two-bedroom apartments in your town. From these 100 samples, 100 confidence intervals were calculated. Based off of our work on sampling distributions, we know that the 100 sample means will be close to the population mean (some might even be the same as the population mean), but some will be closer and some will be farther. Thus some of the confidence intervals will be 'good' estimates of the population mean rent for two-bedroom apartments (that is, the population mean will actually be included in the confidence interval) and some will be 'bad' estimates (that is, the population mean won't actually be included in the confidence interval). Since the population mean is unknown none of the 100 people who made these confidence intervals knows if their estimate is good or bad. Instead, they can only state how confident they are in their estimate. That is, they can only state their level of confidence.

Suppose that all 100 people made 95% confidence intervals. What does that mean? Well suppose a local real estate company has actually worked out the population mean rent for two-bedroom apartments in your town by finding out the rent for all two-bedroom apartments. Since they know the population mean, they don't have to estimate it. They have found it to be $1200.

[link] shows the 100 confidence intervals created by the 100 random samples and compares them to the population mean. If the interval is yellow then that means it is a good estimate. If it is red, then that means it is a bad estimate. The yellow part in the middle represent the 95% confidence interval. The yellow and the blue combined represent the 99% confidence interval.



The above image was created using an applet from

David Lane's onlinestatbook.com[footnote]

Notice that out of the 100 confidence intervals calculated, 93 of them are good estimates (contain $1200) and seven of them are bad estimates (do not contain $1200). This is what the confidence level refers to. That is, if you take many, many random samples of the same size and construct a confidence interval for each of the samples, then the percentage of confidence intervals that contain the population mean is 95% and the percentage that do not contain the population mean is 5%. Thus, the confidence level refers to the probability that the process of creating a confidence interval results in the population parameter being in the confidence interval. It is NOT the probability that the population mean falls in a specific confidence interval. Remember that the population mean is fixed. Therefore, either the population mean does fall in the confidence interval or it doesn't. Since there is no randomness to whether it does fall or not, there is no probability associated with that event. Instead the level of confidence refers to the percent of confidence intervals that contain the parameter being estimated if the study/experiment is repeated many, many times.

What has been described above is not an easy idea. Many people who have studied statistics are under the false impression that the confidence level refers to the probability that the parameter is in the confidence interval. Don't fret if this doesn't make entire sense to you right away. Give yourself some time to think about it and process it.

As a note, the example provided in [link] is a bit surprising. If you flip a fair coin 100 times, you would expect that around 50 heads and 50 tails, but due to sampling variability it would also be fair to get 49 heads and 51 tails. It is the same thing with confidence intervals, we expect that for 100 confidence intervals that around 95 of them contain the population mean and 5 of them don't, but it would be fair to get 94 good estimates and 6 bad ones. Once again, the law of large numbers tells us that as the sample size increases the closer we will get to the 95%. That is, if we take 1000 random samples instead of 100, the more likely it is that 95% will be good estimates and 5% will be bad.

Comparing different levels of confidence for the same random sample

Common choices for confidence levels

The most common choices for confidence levels are 90%, 95%, and 99%, but you can choose the level of confidence to be any percentage between 0.00001% and 99.99999%. The can't choose 100%, because that would mean you for sure know that the population parameter falls within the confidence interval. You also can't choose 0%, because that would mean you for sure know that the population parameter does not fall within the confidence interval. If you knew for sure the parameter falls (or does not fall) in the confidence interval, you wouldn't be bothering to do a confidence interval, because you already know that parameter.

90%, 95%, and 99% are common levels of confidence because they offer a high degree of confidence.

How does the confidence level change the confidence interval? Think about the following two confidence intervals for the mean age of students at your university:

4 years old to 85 years old

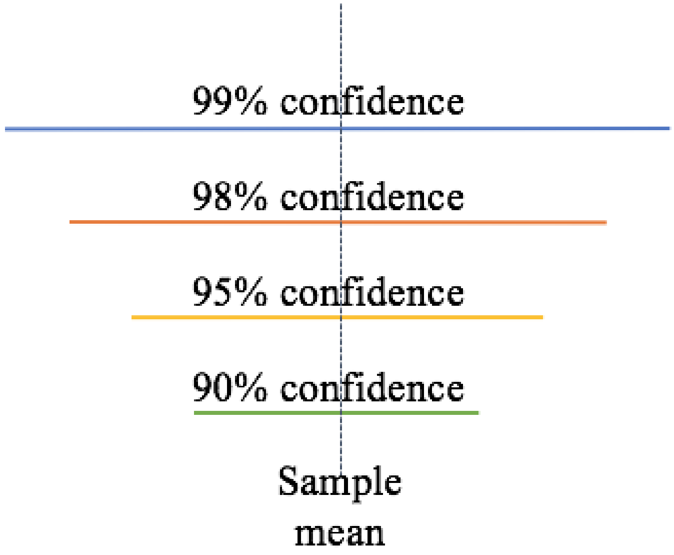
20 years old to 21 years old

Which confidence interval are you more confident actually contains the population mean? Well it is pretty likely that the population mean age of students at your university is somewhere between 4 years old and 85 years old, because the range is so wide that it most likely `catches' the population mean.

In general, the larger the confidence level, the wider the confidence interval. That is, to increase the confidence in the estimate, we make the confidence interval wider so that it is more likely to catch what we are estimating. Think about the confidence interval like a net. The smaller the net, the less likely it is you'll catch the fish. But the wider the net, the more likely it is that you will. Thus for the same sample, the 90% confidence interval is narrower than the 99% confidence interval.

Thus, a 99% confidence interval is very reliable, but it gains reliability at the price of precision. That is, its wideness might come at the sake of usefulness. Going back to the confidence interval for the mean age of students at your university, we can be very confident that the population mean age is between 4 and 85 years old, but that doesn't actually help understand what the population mean age is. We are less confident in the estimate of 20 to 21 years old, but it is providing us more useful information.

To summarize, higher degrees of confidence mean that we are more sure that the parameter fall in the interval (i.e. more reliable). Lower degrees of confidence mean that the interval is smaller and thus gives us a better idea of where the parameter in question is (i.e. more precise). See [link]



The choice of a 95% level of confidence is most common because it provides a good balance between precision and reliability.

What else effects the width of a confidence interval?

The width of the confidence interval is determined by the margin of error, E. In general, the confidence interval is calculated as follows:

point estimate + E, point estimate -E

The size of the margin of error determines the width of the confidence interval. That is, the bigger the margin of error is, the wider the confidence interval.

Factors that effect the size of the confidence interval include the size of the sample, the amount of variability in the data, and the confidence level.

As per the law of large numbers, the larger the sample size, the closer the statistic (or point estimate) is to the parameter. Therefore, the larger the sample size, the less error there is between the statistic and the parameter. This means that **the margin of error is smaller for larger sample sizes** taken from the same population.

The greater the variability in the population, the greater the variability in the statistics. We saw this in Chapter 6 when we determined that the standard deviation of the sampling distribution was related both to the standard deviation of the population and the sample size. That is, the variation between the statistics relied both on the variation in the population and the sample size. Thus, **the margin of error is larger in situations where there is more variability in the population**.

As stated above, the larger the confidence level, the wider the confidence interval. Therefore, **the margin of error is larger for larger levels of confidence**.

Common misconceptions about confidence intervals

1. **The confidence interval contains 95% of the data values.** A confidence interval is an estimate for a parameter (like the population mean or population proportion). Though the data values are used to construct the confidence interval, the confidence interval does not tell us anything about the range of the data values.
2. **We are 95% confident that the sample mean is contained in the confidence interval**. If the confidence interval is for the population mean, then the sample mean has to be in the confidence interval. In fact, it is right in the middle. Remember that the confidence interval for the population mean is calculated as follows: $\bar{x}-E,\bar{x}+E$. All confidence intervals contain the point estimate being used to construct the confidence interval.
3. **Increasing the sample size increases the width of the confidence interval**. In fact, the opposite happens. From the law of large numbers, we know that a larger sample size means that the point estimate will likely be closer to the parameter being estimated. Therefore, as the sample size increases, the margin of error decreases and the width of the confidence interval decreases.
4. **A 90% confidence interval is wider than a**

95% for the same data. Again, it is the opposite that happens. To become more confident in our estimate (i.e. increasing the level of confidence), we widen the confidence interval. A wider confidence interval is a larger net which makes it more likely that we catch the parameter we are estimating.

The Basic Premise of Constructing a Confidence Interval

Overview of how to construct a confidence interval

In the above section, we discussed at length what a confidence interval is. Now we are going to discuss how to construct and interpret one.

A confidence interval is constructed by taking the point estimate and adding and subtracting the margin of error. The margin of error is constructed by looking at the level of confidence and the amount of variation between the point estimates. For example, the margin of error for a confidence interval for a population mean is found by looking at the level of confidence (which the researcher determines) and the amount of variation between the sample means. The amount of variation between the samples means is the amount of variation in the sampling distribution for sample means, i.e. the standard error. Thus **a confidence interval is always constructed from the appropriate sampling distribution**.

This is helpful in two ways:

- From our work in Chapter 6, we know what the standard error is for both the sample mean $\sigma\bar{x}=\sigma n$ and sample proportion $\sigma\hat{p}=\pi(1\text{-}\pi)n$.
- From our work in Chapter 6, we know what the shape of the sampling distribution will be from

the Central Limit Theorem.

The margin of error is found by taking into account the confidence level and the standard error.

The next section examines how the margin of error is constructed for confidence intervals for the mean.

The Confidence Interval Estimate of a Population Mean

Explanation of how to construct a confidence interval for the mean

There are multiple models for finding the confidence interval for the mean. The models we will be looking at rely on the sampling distribution being approximately normal. If that is not the case, then we cannot use these models.

Therefore, the following section relies on the following assumptions:

- The sampling distribution for sample means of the population we are investigating is approximately normally distributed.
    - If the sample size is greater than 30, then the central limit theorem tells us that we can assume that the sampling distribution is approximately normal regardless of the population distribution. Thus, if the sample size is greater than 30, we can use this model.
    - If the sample size is less than 30, the central limit theorem does not guarantee that the sampling distribution of the means will be normal. Therefore, **to use this model the population distribution needs to be approximately normal** so

that we know that the sampling distribution for sample means is normal.

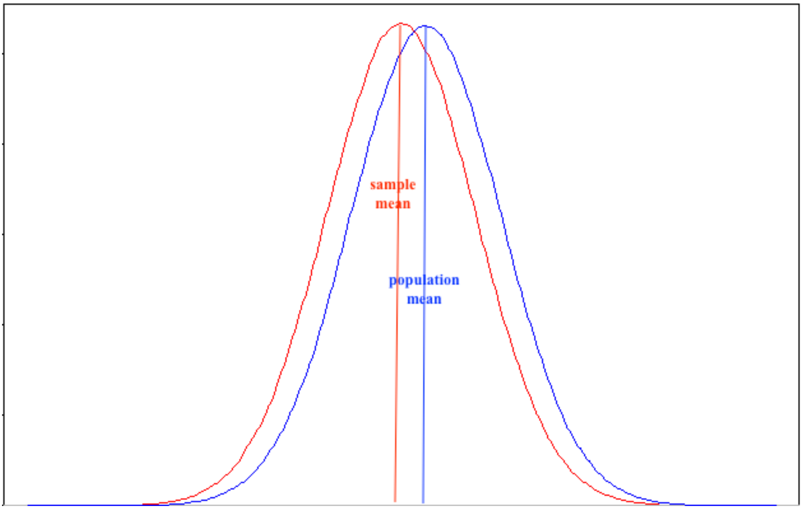
- The sample we are using to construct the confidence interval is a random sample.

To construct a confidence interval for the mean, collect a random sample from the population whose mean is being estimated. Then calculate the sample mean.

The next step is to calculate the margin of error. To do this, we begin by finding out how much sampling variability there is in the sampling distribution. That is, we determine how much variation we expect between the sample means. This is found by calculating the standard error of the sampling distribution for sample means:
$\sigma \bar{X} = \sigma \bar{X} n$

Now we want to take into account the level of confidence. To do this, we construct a normal distribution that is centred at the sample mean, $\bar{x}$, whose standard deviation is the standard error of the mean, $\sigma\bar{X}n$. The data values for this distribution are sample means. Therefore this is a sampling distribution for sample means. This sampling distribution is an estimate of what the sampling distribution of the population will look like:
Blue curve: True sampling distribution for sample means centred at μx and with a standard deviation

of $\sigma\bar{X}n$. **Red curve**: Estimate of the true sampling distribution for sample means based on the mean of the random sample. It is centred at $\bar{x}$ and has a standard deviation of $\sigma\bar{X}n$.

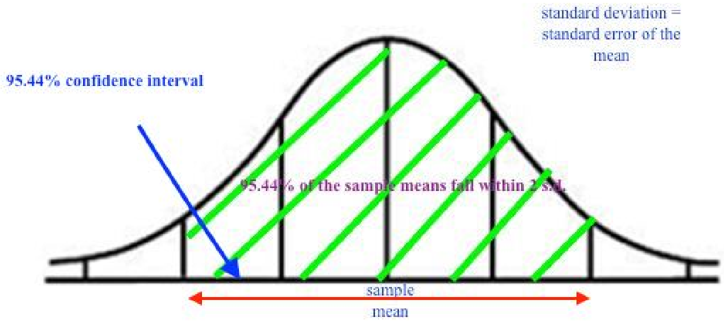


In [link], the blue sampling distribution is the theoretical sampling distribution of the population, which is unknown. The red sampling distribution is an estimate of the blue curve based on the sample mean found from the random sample. We will use the red sampling distribution to estimate the population mean.

Using the red sampling distribution, we want to determine the interval of sample means that fall within a specific percentage from the mean. The specific percentage is the confidence level.

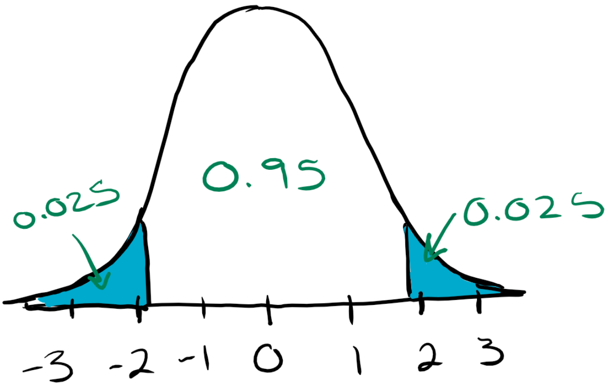
Suppose that the confidence level is 95.44%. From the empirical rule, we know that 95.44% of data

values fall within 2 standard deviations of the mean for normally distributed data. Therefore, if we wanted to construct a 95.44% confidence interval, we would take the sample mean and add and subtract two standard deviations from it. Since we are dealing with a sampling distribution, the standard deviation we are referring to is the standard error of the mean. Therefore, a 95.44% confidence interval is found by calculating $\bar{X} \pm 2\cdot\sigma\bar{X} = \bar{X} \pm 2\cdot\sigma Xn$. Thus for a 95.44% confidence interval, the margin of error is $E = 2\cdot\sigma Xn$.
95.44% confidence interval for the mean


standard deviation =
standard error of the
mean
95.44% confidence interval
95.44% of the sample means fall within 2 s.d.
sample
mean

If we wanted to find a 95% confidence interval, we would use the same process, but we would want a slightly narrower interval. Therefore, instead of multiplying the standard error by 2, we would multiply it by a slightly smaller number. To determine by what number, we would need to find out how many standard deviations away from the mean results in an area of 95%. In other words, we would need to find the z-score that gives an area of 95%.

Standard normal curve with the area of the tails being 5%.



If the area in the middle of the curve is 95%, then the area of one tail is 2.5%. Using a computer program, we can find this value to be ±1.96.

To do this, go to your computer program and go to the menu option that lets you find probabilities for normal distributions. Then make the mean 0 and the standard deviation 1. Then switch from calculating probabilities to finding z-values (like you are going to find a percentile). In the appropriate box, put 0.0025 in for the area in the upper tail. When you hit enter, the program will give you 1.96 as the z-value for this area.

In general, the value that you multiply the standard error by is called the **critical value** and is denoted by $z\alpha/2$, where α is the total area of the tails. $(1-\alpha)\times 100\%$ is the level of confidence.

The margin of error is $E=z\alpha/2\times\sigma Xn$

The confidence interval is $x^{-}\pm E$. As it is an interval, always write it with the smaller number first (x^{-}-E) followed by the larger number ($x^{-}+E$).

Suppose that a random sample of 175 students from a university is taken and their average age is 21.34 years old and the population standard deviation is known to be 5.12 years.

1. Find the 95% confidence interval for the population mean age of all university students.
2. Interpret the confidence interval in the context of the question.
3. Explain what the level of confidence means in the context of the problem.
4. If we decreased the sample size to 100, what would you expect to happen to the confidence interval? Explain your answer.
5. Suppose that an administrator at the university claims that this university caters to older students and that the mean age is 23. Does the confidence interval support the claim?

---

1. We can use the standard normal model to find the confidence interval, because the sample was collected randomly and, since

the sample size is greater than 30 (it is 175), we can be very confident that the sampling distribution for the sample means is normal due to the central limit theorem. To find the confidence interval, use a computer program. Make sure to choose the z-model (instead of the t-model). Input the sample size as 175, the sample mean as 21.34 and the standard deviation as 5.12. Choose the level of confidence to be 95%. This gives the following output:

95%	confidence level
1.96	z
0.759	margin of error
20.581	lower confidence limit
22.099	upper confidence limit

From this, we can see that the confidence interval for the mean is 20.58 to 22.10.

2. To interpret the confidence interval, we would say that we are 95% confident that the population mean age of students from this university is somewhere between 20.58 years old and 22.10 years old. That is, we are estimating that the population mean age is somewhere between 20.58 years old and 22.10 years old.
3. The confidence level means that if we took many random samples of size 175 from the

student body of this university and constructed many confidence intervals for each of these random samples, then 95% of these confidence intervals will contain the population mean age for this university, while 5% will not.

4. If the sample size is decreased to 100, we would expect that the confidence interval would get wider. From the law of large numbers, we know there is more sampling variability in smaller samples. Thus there is more potential for error between the sample mean and the population mean when the sample size is smaller. The margin of error then is bigger to take this into account. This is supported by the formula for the margin of error ($z\alpha/2 \times \sigma n$). Since we are dividing by the n, the margin of error would be smaller for larger n and bigger for smaller n.
5. We have estimated that the population mean age is between 20.58 years old and 22.10 years old. Therefore, based on our estimate, it is unlikely that the mean age of this university is 23 years old as 23 does not fall within our estimate. The administrator's claim is most likely incorrect.

A few notes about the above confidence interval:

- All of the means in the interval are equally likely. That is, each of the estimates of the population mean in the interval have an equal chance of being correct. For example, 20.58 years old and 21.25 years old are both equally likely estimates of the population mean age.
- The sample mean of 21.34 is right in the middle of the interval.
- The margin of error is 0.759 and is found using the formula $z\alpha/2 \times sn = 1.96 \times 5.12175$
- It is possible that the population mean is not captured by this confidence interval, but we wouldn't know whether it does or not without knowing the population mean.

Wait a second! If we don't the population mean (μx), how do we know the population standard deviation (σx) in the standard error formula???

That's a really good question. The actual formula for the population standard deviation involves knowing the population mean: $\sigma x = \Sigma(X\text{-}\mu)2n$. Therefore, if we don't know the population mean, how do we know the population standard deviation?

There are two possible answers to this:

1. In some long running process (e.g. manufacturing), the standard deviation may be very static. Therefore, the population standard deviation could be known even if the population mean isn't.
2. We don't know the population standard deviation, so instead we estimate it with the sample standard deviation.

It is fairly unlikely that in most situations, the population standard deviation will be known. Thus, we will focus on situations where the population standard deviation is unknown. In that case, we will use the sample standard deviation s to estimate the population standard deviation σx.
The Student-t distribution was created by William Gosset, an English statistician who worked for Guinness breweries. While working for Guinness, Gosset developed the Student-t distribution, but was prohibited from publishing his work by his employers who worried about trade secrets getting out. Thus he published his work under the pseudonym `Student' in 1907. The distribution, then, should really be called the Gosset-t distribution. Comparison of Student-t distribution with standard normal distribution Critical value for Student-t distribution with n=5

Student-t distribution

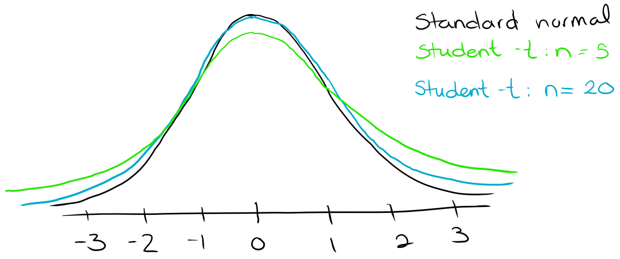
To use this model to construct a confidence interval, we need to again assume that the sampling distribution is normal and that the sample was collected randomly. Just as we saw above, there are two general situations that need to occur to ensure the sampling distribution is normal:

- If the sample size is greater than 30, then the central limit theorem tells us that we can assume that the sampling distribution is approximately normal regardless of the population distribution. Thus, if the sample size is greater than 30, we can use this model.
- If the sample size is less than 30, the central limit theorem does not guarantee that the sampling distribution of the means will be normal. Therefore, **to use this model the population distribution needs to be approximately normal** so that we know that the sampling distribution for sample means is normal.

Since we don't know the population standard deviation, we will be using the sample standard deviation to estimate σx. That means we are estimating the population mean using the sample mean and sample standard deviation. This suggests that there may be more error in our estimate. To account for the greater error, we want the confidence interval to be slightly wider. To do this the margin of error needs to slightly bigger. The margin of error is the critical value × the standard error. The standard error is inherent to the population and can't be changed, but the critical value can be. So instead of using the standard normal distribution to find the critical value, we use the Student-t distribution [footnote]

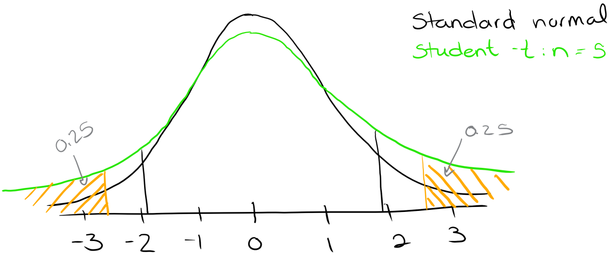
Here is some information about the Student-t distribution.

- The Student-t distribution is a normal distribution with $\mu = 0$ and $\sigma > 1$. The standard deviation of the Student t distribution is different for different sample size. Remember that the standard normal distribution is a normal distribution with $\mu = 0$ and $\sigma = 1$. Therefore, the Student-t distribution is centred at the same place as the standard normal distribution, but has greater variation so it is slightly wider and shorter. See [link].
- The smaller the sample size, the greater the variability is in the sampling distribution. When the sample size is larger, there is less variability in the sampling distribution. These aspects are reflected in shape of the Student-t distribution.
- As the sample size n gets larger, the Student-t distribution gets closer to the standard normal distribution.



The standard deviation of the Student-t distribution is based on the **degrees of freedom**, which in turn are based on the sample size. The number of degrees of freedom for a sample corresponds to the number of data values that can vary after certain restrictions have been imposed on all data values. Another way of saying it, is the degrees of freedom are the number of components that need to be known before a statistic is entirely determined. Depending on the model used, the degrees of freedom have a different formula. For this model (i.e. confidence interval for one population mean), the degrees of freedom are the sample size minus 1, i.e. n-1.

As stated above, we want the width of the confidence interval to be wider to take into account the larger variation due to the estimate of the standard deviation. As you can see from the figure above, the Student-t distribution is wider than the standard normal distribution. Which means that the critical value for a 95% confidence level will be greater than that for the standard normal. See the image below.



Notice the critical value is happening about halfway between ± 2 and ± 3. But the critical value for the standard normal distribution is ± 1.96.

The margin of error for this model is:
$E = t \alpha / 2 \times s n$

The confidence interval is constructed in the same way: $\bar{x} \pm E$.

> A manufacturer of AAA batteries wants to estimate the mean life expectancy of the batteries. It is known that the life expectancy of such batteries is typically normally distributed.
>
> A random sample of 25 batteries has a mean of 44.25 hours and a standard deviation of 2.25

hours. Assume the population is normal.

1. Construct a 95% confidence interval for the mean life expectancy of all the AAA batteries made by this manufacturer.
2. Interpret the 95% confidence interval.
3. If the confidence level is decreased to 90%, how does the confidence interval change?

---

1. We can use the Student-t distribution model to construct the confidence interval, because the population standard deviation is unknown (so we don't use the standard normal distribution), the sample is collected randomly, and the sampling distribution of the sample means is normal because the population distribution is normal. To find the confidence interval, use a computer program. Make sure to choose the t-model (instead of the z-model). Input the sample size as 25, the sample mean as 44.25 and the standard deviation as 2.25. Choose the level of confidence to be 95%. This gives the following output:

95%	confidence level
2.064	t
24	degrees of freedom
0.929	margin of error
43.321	lower confidence

	limit
45.179	upper confidence limit

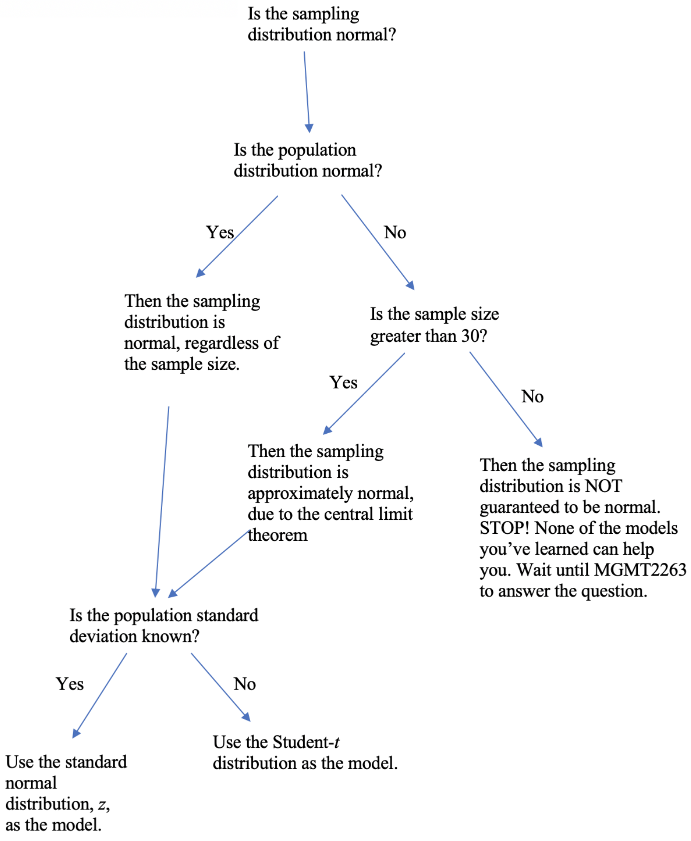
From this, we can see that the confidence interval for the mean is 43.321 to 45.179.

2. To interpret the confidence interval, we would say that we are 95% confident that the true mean battery life of brand of AAA batteries is somewhere between 43.32 hours and 45.18 hours.
3. If the confidence level is decreased to 90%, we would expect that the confidence interval would get narrower. A higher level of confidence is obtained by making the confidence interval wider. Therefore, if the confidence level is decreased, then the confidence interval would get narrower.

Notice from the computer output, that the critical value is 2.064 with 24 degrees of freedom (i.e one less than the sample size). If the population standard deviation was known, the critical value would be 1.96. To re-iterate, since we are estimating the population standard deviation with the sample standard deviation, we know there is more room for error in the estimate. Therefore, we want the estimate (i.e. confidence interval) to be slightly wider, thus the margin of error needs to be slightly bigger. This is done by using the Student-t distribution, which results in bigger critical values for the same confidence level as would occur for the standard normal distribution. In this case, 2.064.

[link] is a flow chart that indicates how to make a choice of which model to use to construct a confidence interval (CI) for the mean.

Flow chart for determining which model to use when constructing confidence interval for the mean


Is the sampling distribution normal?
Is the population distribution normal?
Yes
No
Then the sampling distribution is normal, regardless of the sample size.
Is the sample size greater than 30?
Yes
No
Then the sampling distribution is approximately normal, due to the central limit theorem
Then the sampling distribution is NOT guaranteed to be normal. STOP! None of the models you've learned can help you. Wait until MGMT2263 to answer the question.
Is the population standard deviation known?
Yes
No
Use the standard normal distribution, z, as the model.
Use the Student-t distribution as the model.

Sample Size Determination

Determining an appropriate sample size is very important. Too small of a sample may lead to poor results. Too large of a sample needlessly wastes time and money.

Prior to this section, we would have determined if a sample size was large enough simply by guessing. Here we will learn a formula for finding the appropriate sample size based on the amount of error we will accept in our results. This can be done by determining the minimum sample size needed to have a certain margin of error. To do this, we solve for the sample size n in the margin of error formula. $E = z \alpha 2 \cdot s n n = z \alpha / 2 \cdot s E n = z \alpha / 2 \cdot s E 2$

As we would always rather than have one more object of study rather than one less, we will always round up the result of this calculation. That is, if the result of the formula is 50.2, then we will round up to 51.

A couple of notes about the formula:

1. Since n is unknown we can't use t. Think about why this is so.
2. We still need to have a sense of the standard deviation to use this formula. As such, we will often do a preliminary study to estimate of the standard deviation.

You plan to do a study of hypnotherapy to determine how effective it is in increasing the number of hours of sleep participants get each night. To do this you will measure the number of hours of sleep for each of the participants after they've done hypnotherapy. You want to ensure that your estimate for the mean number of hours of sleep is within 0.2 hours of the true mean with a 95% level of confidence. Prior to doing the full study, you do a pilot study with 12 participants, which provides the following data:
$8.2; 9.1; 7.7; 8.6; 6.9; 11.2; 10.1; 9.9; 8.9; 9.2; 7.5; 10.5$

How many participants should be in your study?

We know the confidence level (95%). The margin of error is stated by saying that we want the estimate of the true mean to be within 0.2 hours. Thus the 0.2 hours is telling us how much error we want in the estimate (i.e. $E = 0.2$). We do need to have a sense of the standard deviation, which we get from the preliminary study. Using the 12 participants, we get a sample standard deviation of 1.29.

We can now use a computer program to do the calculation. From the question, we know the margin of error (E) is 0.2, the standard deviation is 1.29, and the confidence level is 95%. When we input this into the computer program, we get output similar to this.

95%	confidence level
1.96	z
159.814	sample size
160	rounded up

From this, we can see that to get our sample size within 0.2 hours of the true mean we would need a sample size of at least 160 participants.

The Confidence Interval Estimate of a Population Proportion

Explanation of how to find and interpret a confidence interval for proportion and sample size determination.

Here we want to construct a confidence interval to estimate the population proportion π based off of the point estimate of the sample proportion $\hat{p}$.

Confidence intervals for proportion are constructed by taking the point estimate $\hat{p}$ and adding and subtracting the margin of error E: $\hat{p} \pm E$.

There is more than one model for constructing a confidence interval for the sample proportion. The model we will discuss here has the following criteria:

- The variable being studied satisfies the conditions of the binomial distribution.
- The sampling distribution for sample proportions is approximately normal. This occurs if the number of successes $(n \times \pi)$ is at least 5 and the number of failures $(n \times (1-\pi))$ is at least 5. As π is unknown this can be checked by determining if the number of successes and failures in the sample are both at least 5.

The margin of error is found in a similar way to margin of error for the mean. That is, it is the critical value × the standard error. As we are assuming that the sampling distribution is approximately normal, we will use the standard normal distribution to find the critical value. Since the variable being studied satisfies the conditions of the binomial distribution, we know from Chapter 6 that the standard error of the sampling distribution is π(1-π)n. As we don't know π as that is what we are trying to estimate, we will estimate π in the formula with the sample proportion pˆ. This results in the estimate of the standard error to be pˆ(1-pˆ)n

If these conditions are met, then the formula for the margin of error is:
$E = z \alpha / 2 \times p \hat{} (1 - p \hat{}) n$

Example: Cell phones

Suppose that a market research firm is hired to estimate the percent of adults living in a Vancouver who have cell phones. Five hundred randomly selected adult residents in Vancouver are surveyed to determine whether they have cell phones. Of the 500 people sampled, 421 responded yes - they own cell phones.

1. Using a 92% confidence level, compute a confidence interval estimate for the true proportion of adult residents of this city who have cell phones.
2. Would it be appropriate to say that 85% of

residents have a cell phone in Vancouver?
3. What does the confidence level tell us in the context of the question?

Solutions:

1. We can use the standard normal model for proportions to construct our confidence interval as the variable (cell phone ownership) follows a binomial distribution (1: The variable is random (random sample); 2: The outcomes are being counted (number of people who have cell phones); 3: There is a fixed number of trials (500); 4: There are two possible outcomes (have cell phone or don't have cell phone); 5: Though π is unknown it is fair to assume that the proportion of people who have a cell phone on a given day in Vancouver is very stable) and the sampling distribution for proportions is normal as the number of successes is 421 and the number of failures is 79 (i.e. they are both greater than 5). Use a computer program to construct the confidence interval. Input x as 421 (this may be in the same place as the sample proportion, but when you input the whole number it will switch to x), the sample size as 500, and the confidence level as 92%. Notice that you don't have to state whether it is z or t as there is only one model for this situation. This gives the following output:

92%		confidence level

1.751	z	
0.029	margin of error	
0.813	lower confidence limit	
0.871	upper confidence limit	

From this, we can see that the confidence interval for the mean is 0.813 to 0.871.

2. To interpret the confidence interval, we would say that we are 92% confident that proportion of residents of Vancouver that own a cell phone is somewhere between 81.3% and 87.1%.
3. Since 85% is contained in the confidence interval, it is appropriate to say that the proportion of residents in Vancouver who have a cell phone is 85%.
4. The confidence level means that if we took many random samples of Vancouver residents of size 500 and constructed many confidence intervals for each of these random samples, then 92% of these confidence intervals will contain the population proportion of cell phone users, while 8% will not.

A couple of notes about the confidence interval:

- The margin of error is 0.029 or 2.9%. The margin of error for a confidence interval for proportions has to be less 1 (or 100%). If the sample size is large enough, the margin of error should be quite small (less than 10%).
- Since proportions can only range from 0 to 1 or

0% to 100%, the confidence interval can never exceed these values. For example, if the sample proportion is 92% and the margin of error is 10%, then the confidence interval would be 82% to 102%, but since the upper bound is impossible, we would round the answer to 82% to 100%.

Determining sample size

Just like with the mean, we want to determine an appropriate sample size to achieve a maximum amount of error in our estimate for the population proportion.

To find the formula for n, we again solve for n in the formula for the margin of error, this results in the following formula:
n = z α / 2 2 p ˆ (1 - p ˆ) E 2

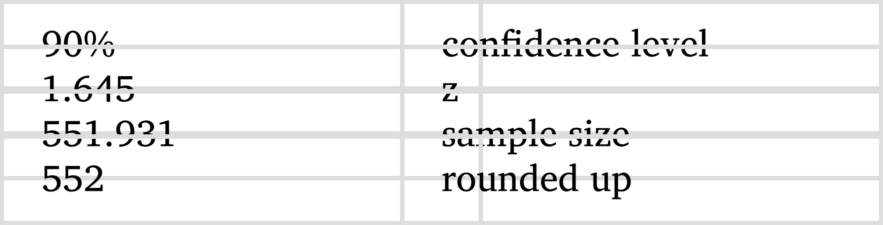
To use this formula we need to know the margin of error, the confidence level and the sample proportion.

Note: If no estimate for π exists, then use pˆ=0.5.

The Western Canada Communications Company is considering a bid to provide long-distance phone service. You are asked to conduct a poll to estimate the percentage of consumers who are satisfied with their current long-distance phone service. You want to be 90% confident that your sample percentage is within 2.5 percentage points of the true population value, and a Roper poll suggests that this percentage should be about 85%. How large must your sample be?

The confidence level is 90%, the sample proportion is 85%, and the amount of error we want in our estimate (i.e. the margin of error) is 2.5%.

We can now use a computer program to do the calculation. From the question, we know the margin of error (E) is 0.025 (remember to write it as a decimal), the sample proportion is 0.85, and the confidence level is 90%. When we input this into the computer program, we get output similar to this.



90%	confidence level	
1.645	z	
551.931	sample size	
552	rounded up	

From this, we can see that we need to have at least 552 consumers in our sample.

Introduction to One Population Hypothesis Testing
This section offers a summary of the general concept and purpose of a hypothesis test. The section discusses how a sample statistic must be examined in order to investigate whether the value of a population parameter has changed from what has previously been claimed or believed. The concept of likely and unlikely observations under the assumption of a prevailing claim is explained.

What are hypothesis tests?

In chapter 7, you learned how to construct an estimate of a population parameter, such as a mean or proportion, from a sample statistic. In this chapter we examine a related concept: investigating whether the value of a population parameter has changed from what has previously been claimed or believed. Again, we use the sample data for this investigation.

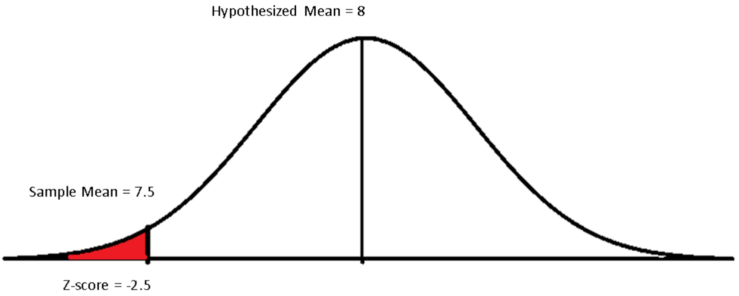
For example, it is commonly stated that adults should get 8 hours of sleep per night. Many of us may suspect that the real average is lower. In conducting an investigation, since we don’t yet have evidence to the contrary, we will treat the mean of 8 as the prevailing claim. In other words, we must assume the true population mean is 8 unless we can prove otherwise. In our attempt to find proof against the prevailing claim, we would need to gather sample evidence.

Let's say that after gathering a large random sample (say, $n = 50$), you discover that the sample mean number of hours slept per night is only 7.5. So is a sample mean of 7.5 hours *proof* that the true population mean is not 8 hours, as claimed, but actually less? On the surface, it would appear so. However, recall from chapter 7 that every sample mean will be different from the true population mean. Some sample means will be a little different and others will be very different.

Also, recall that *all possible* sample means taken from a population, plotted on a distribution, is called a sampling distribution of sample means. The mean or middle of this distribution will be the true population mean, which at present we are assuming to be 8. And if 8 really is the true population mean, then most sample means would be expected to be very close to 8, but some--those means near the tails of the distribution--could be much lower or much higher than 8. The figure below shows a normal curve with a mean of 8 and a standard error of 0.20. As the curve expands towards the tails, the number of observations we would expect to see gets smaller and smaller. In other words, sample means that come from far out in the tails of the distribution are considered rare or unlikely occurrences. So for this example, the question is whether 7.5 is so far out into one of the tails that it would be considered an unlikely observation under the assumption that the middle of this curve is actually 8.

To measure how far into the tail our sample mean of 7.5 is, we must use a familiar measuring tool called a Z score (or a T score for smaller samples). Since we are assuming the mean or middle of our sampling distribution is 8 (remember that 8 is our prevailing claim), we need to measure the number of Z scores our sample mean of 7.5 is from 8. Recall from chapter 6 that a variable's Z score is simply the number of standard deviations the variable lies from the middle of the normal curve. Also recall that over 95% of a normally shaped distribution will fall within two Z scores (two standard deviations) of the middle and over 99% will fall within three Z scores.

In hypothesis testing, any value falling more than two standard deviations from the middle would be considered *unlikely* (less than 5% of all possible sample means will fall more than two standard deviations from the middle). Any value falling more than three standard deviations from the middle would be considered very unlikely (less than 0.5% of all possible sample means will fall more than three standard deviations from the middle). If the standard error for our example is 0.2, then our sample has a Z score of -2.5 (7.5 – 8/0.2). That is, our sample mean of 7.5 lies 2.5 standard deviations to the left of our hypothesized population mean, well out into the left tail of the curve. So, it does appear that our sample mean can be considered an unlikely occurrence. The conclusion then must be that if the true population mean is actually 8 it would unlikely for us to obtain a sample mean as small as 7.5. But since we *did* obtain such a mean, we must therefore conclude the true population mean is less than 8.



Without getting into further technicalities at this point, we have shown that a hypothesis test seeks to measure whether the sample evidence can be considered unlikely under the assumption that the prevailing claim is true. If our answer is 'yes', then we have good reason to reject the prevailing claim. If our answer is no, then we must let the prevailing claim stand, at least until stronger evidence against it is found.

In the next section we will break down the various steps in a hypothesis test.

The Distribution Needed for Hypothesis Testing

In chapter 6, we discussed sampling distributions, which are used for hypothesis testing. We will perform hypotheses tests of a population mean using two particular sampling distributions: a normal distribution or a Student's *t*-distribution. We will perform hypothesis tests of a population proportion using a normal sampling distribution that has been approximated from a binomial situation.

Central Limit Theorem Revisited

When you perform a **hypothesis test of a single population mean *μ*** using a normal distribution (often called a *z*-test), you take a large random sample from the population. When working with large samples, you should recall from chapter 6 that Central Limit Theorem says that the sampling distribution of means will be approximately normal even if the population from whence the sample came is not. For this reason we can perform hypothesis tests using large samples and the normal distribution regardless of the shape of the parent population.

Many statisticians prefer to use a t-distribution if the population standard deviation is unknown, even if the samples are large. The reasoning behind this is that using the sample standard deviation in place of the unknown population standard deviation adds an extra degree of potential error that can only be accounted for by using a t- distribution. However, as noted in the previous chapter, it is common practice to use the normal (Z-based) sampling distribution when working with large samples. Specifically, when $n > 40$, we will use the standard normal(z-based)distribution to conduct a hypothesis test..

When working with small samples, we will perform a **hypothesis test of a single population mean** μ using a **Student's *t*-distribution** (often called a t-test). There are fundamental assumptions that need to be met in order for the test to be considered valid. Most importantly, since Central Limit Theorem does not apply to small samples, we have no guarantee the the sampling distribution will be normally shaped. For this reason, we can only perform means tests with small samples when we know the population is normally distributed.

Please see Figure 6 in the previous chapter for further insight into how to determine which sampling distribution is appropriate when conducting a hypotheses test of a population mean.

When you perform a **hypothesis test of a single population proportion *p***, you take a random sample from the population. You must meet the conditions for a **binomial distribution** which are: there are a certain number *n* of independent trials, the outcomes of any trial are success or failure, and each trial has the same probability of a success *p*. The Central Limit Theorem says the shape of the binomial distribution will approximate the shape of the normal distribution if the sample is sufficiently large. To ensure this, the quantities *np* and *nq* must both be greater than five ($np > 5$ and $nq > 5$). Then the binomial distribution of a sample (estimated) proportion can be approximated by the normal distribution with $\mu = p$ and σ= pq n . Remember that $q = 1 - p$.

Large Sample Hypothesis Test for the Mean

Going back to the standardizing formula we can derive the **test statistic** for testing hypotheses concerning means. We have already worked with the formula below when introduced to sampling distributions in Chapter 6. You should, however, notice one small difference. When we perform hypothesis tests, we don't know the population mean; we simply have a claim or belief about the mean, which may or may not be true. Because the mean is hypothesized rather than known, we use a

slightly different symbol in the equation, μ_0, as seen below.
$Zc = x\text{–-}\mu 0\ \sigma/n$

This calculated Z is nothing more than the number of standard deviations that the sample mean is from the **hypothesized** population mean. If the sample mean falls "too many" standard deviations from the hypothesized mean we conclude that the **sample** mean is **unlikely** to have come from a distribution centred around the hypothesized mean.

So how do we know if a sample mean can be considered to have fallen "too many" standard deviations away from a hypothesized mean? Obviously, we can't simply make this decision arbitrarily. Thankfully, we have already been introduced this concept when we examined confidence intervals in the previous chapter. Just as we predetermine our level of confidence before we compute an estimate of a population parameter, so too must we predetermine how strong we need our sample evidence to be (i.e. how many standard deviations away from the hypothesized population parameter it must lie) before we would be confident in rejecting the null hypothesis. This predetermined level in hypothesis testing is called the level of significance, and it is simply 1- the level of confidence. The level of significance is denoted as alpha (α).

This level of significance delineates a set number of standard deviations between evidence that would be considered unlikely and evidence that would be considered not unlikely under the assumption that the null hypothesis is true. By way of example, say we set our level of significance at 5%. The corresponding Z score for a 5% level of significance is 1.645. This means that if our sample mean falls more than 1.645 standard deviations away from the hypothesized middle of the distribution (i.e. the null hypothesis), we can conclude the sample evidence is strong enough to be considered an unlikely event and we can therefore reject the null hypothesis.

Before proceeding further, it's worth reviewing this notion of a significance level from another perspective. The significance level can be thought of as the allowable amount of error in our test. Just as a 95% confidence level will produce an incorrect estimate 5% of the time, so will our hypothesis test with a level of significance set at 5%, produce an incorrect conclusion 5% of the time, at least theoretically. When we set the significance level at, say 5%, we are essentially saying that on our sampling distribution any sample mean that falls into the top (or bottom) 5% of the tail would be considered strong evidence against the null hypothesis. This does not mean the evidence is perfect, however. There is certainly the possibility that a sample mean that falls into the top (or bottom) 5% of the tail could have come from a population in which the null hypothesis is true. Indeed that possibility is actually 5%. But 5% is a pretty small number, which is why we would say the observance of such a sample mean must be considered an unlikely--but not impossible-- event.

Small Sample Hypothesis Tests for the Mean

Because the samples are small and we don't know the population standard deviation, we must use a Student t-distribution rather than a Z distribution to perform our tests. The new standardizing formula below will be used to compute how many standard deviations our sample mean falls from the hypothesized middle of the t-distribution.

tc= X– -μ0 s/n

Large Sample Tests for the Proportion

When conducting a hypothesis test on a proportion, we can use a Z-based test so long as the sample is sufficiently large. A sample is considered large if npˆ and n(1-pˆ) both exceed 5. Even though we will perform a Z-based test, because we are working with proportions, the standardizing formula is quite different. In the numerator, the hypothesized proportion is subtracted from the observed sample proportion. In the denominator, the standard error is calculated by first multiplying the hypothesized proportion by 1 - the hypothesized proportion; then by dividing the result; and finally taking the square root of that result.

$Z^* = \hat{p} - \pi o/\pi o(1-\pi o)n.$

Chapter Review

In order for a hypothesis test’s results to be generalized to a population, certain requirements must be satisfied.

When testing for a single population mean:

1. A Student's *t*-test should be used if the data come from a small, random sample and the population is approximately normally distributed.
2. The normal z-test can be used if the data come from a large, random sample. The population does not need to be normally distributed.

When testing a single population proportion use a normal test for a single population proportion if the data comes from a random sample, fit the requirements for a binomial distribution, and the mean number of success and the mean number of failures satisfy the conditions: $np > 5$ and $nq > n$ where n is the sample size, p is the probability of a success, and *q* is the probability of a failure.

Which two distributions can you use in hypothesis testing for the mean in this chapter?

A normal distribution or a Student's *t*-distribution

Which distribution do you use when the sample size is small, the standard deviation is not known and you are testing one population mean?

Use a Student's *t*-distribution

A population has a mean is 25 and a standard deviation of five. The sample mean is 24, and the sample size is 108. What distribution should you use to perform a hypothesis test?

a normal distribution for a single population mean

You are performing a hypothesis test of a single population mean using a Student's *t*-

distribution. What must you assume about the distribution of the data?

It must be approximately normally distributed.

You are performing a hypothesis test of a single population proportion. What must be true about the quantities of npˆ and n(1-pˆ)

They must both be greater than five.

You are performing a hypothesis test of a single population proportion. The data come from which distribution?

binomial distribution

Homework

It is believed that Medicine Hat Community College (MHCC) Intermediate Accounting students get more than seven hours of sleep per night, on average. A survey of 22 MHCC Intermediate Accounting students generated a mean of 7.24 hours with a standard deviation of 1.93 hours. At a level of significance of 5%, do MHCC Intermediate Accounting students get more than seven hours of sleep per night, on average? The distribution to be used for this test is X – ~ ___________

1. Z(7.24, 1.93 22)
2. Z(7.24,1.93)
3. t_{22} df
4. t_{21} df

---

d

Glossary

Binomial Distribution

a discrete random variable (RV) that arises from Bernoulli trials. There are a fixed number, *n,* of independent trials. “Independent” means that the result of any trial (for example, trial 1) does not affect the results of the following trials, and all trials are conducted under the same conditions. Under these circumstances the binomial RV X is defined as the number of successes in *n* trials. The notation is: $X \sim B(n, p)$ $\mu = np$ and the standard deviation is σ= npq . The probability of exactly *x* successes in *n* trials is

$P(X=x)=(n x) p x q n-x$.

Normal Distribution

a continuous random variable (RV) with pdf $f(x)= 1 σ 2π e - (x-μ) 2 2 σ 2$, where μ is the mean of the distribution, and σ is the standard deviation, notation: $X \sim N(\mu, \sigma)$. If $\mu = 0$ and $\sigma = 1$, the RV is called **the standard normal distribution**.

Standard Deviation

a number that is equal to the square root of the variance and measures how far data values are from their mean; notation: *s* for sample standard deviation and *σ* for population standard deviation.

Student's *t*-Distribution

investigated and reported by William S. Gossett in 1908 and published under the pseudonym Student. The major characteristics of the random variable (RV) are:

- It is continuous and assumes any real values.
- The pdf is symmetrical about its mean of zero. However, it is more spread out and flatter at the apex than the normal distribution.
- It approaches the standard normal distribution as *n* gets larger.
- There is a "family" of t distributions:

every representative of the family is completely defined by the number of degrees of freedom which is one less than the number of data items.

Test Statistic

The formula that counts the number of standard deviations on the relevant distribution that estimated parameter is away from the hypothesized value.

Critical Value

The *t* or *Z* value set by the researcher that measures the probability of a Type I error, α.

The Null and Alternative Hypotheses

The actual test begins by considering two **hypotheses**. They are called the **null hypothesis** and the **alternative hypothesis**. These hypotheses contain opposing viewpoints.

H0: **The null hypothesis:** The null hypothesis is the opposite of what the researcher is trying to show. It is the assumption made about a population parameter, such as the mean or proportion. It is a statement that we will assume to be true until we can find strong evidence to the contrary. You can think of the null hypothesis as the assumption that nothing has changed, nothing is different. If you find evidence that suggests the assumption is not valid, then you will reject the assumption about the population parameter in favour of a claim. If you do not find enough evidence that suggests the assumption is not valid, then you do not have enough evidence to support the claim, but that does not mean the assumption is valid.

Ha: **The alternative hypothesis:** This is the claim about the population that the researcher is trying to show and it is contradictory to H0 . It is what we conclude to be likely to be true if our sample evidence suggests that H0 is no longer valid. The alternative hypothesis says that something is different, that things have changed. It must be supported by significant evidence to overthrow the assumption.

Since the null and alternative hypotheses are contradictory, you must examine evidence to decide if you have enough evidence to reject the null hypothesis or not. Since we rarely have access to population data, we must take our evidence from sample data.

Later we will discuss in more detail how to determine if the sample evidence can be considered strong enough to support the alternative hypothesis. Once you have examined the sample evidence, you can determine if it supports the alternative hypothesis or not and make your final **decision.** There are two options for this decision. They are "reject *H0*" if the sample information favours the alternative hypothesis or "fail to reject *H0*" or "decline to reject *H0*" if the sample information is insufficient to reject the null hypothesis. These conclusions are all based upon a level of significance that is set by the analyst.

Table 9.1 presents the various hypotheses in the relevant pairs. For example, if the null hypothesis is equal to some value, the alternative has to be not equal to that value.

H_0	***H_a***
equal (=)	not equal (≠)
greater than or equal to (≥)	less than (<)
less than or equal to (≤)	more than (>)

Note

As a mathematical convention H_0 always has a symbol with an equal sign in it. H_a never has a symbol with an equal in it. The choice of symbol depends on the wording of the hypothesis test.

H_0: The average amount of sleep adult Canadians get per night is greater than or equal to 8 hours.
H_a: The average amount of sleep adult Canadians get per night is less than 8 hours.
H_0: $\mu \geq 8$
H_a: $\mu < 8$

We want to test whether the mean GPA of students in Canadian universities is different from 2.0 (out of 4.0). The null and alternative hypotheses are:
H_0: $\mu = 2.0$
H_a: $\mu \neq 2.0$

We want to test if university students take more than four years to graduate from university, on the average. The null and alternative hypotheses are:
H_0: $\mu \leq 4$
H_a: $\mu > 4$

We want to test if the proportion of Liberal supporters has dropped since the election.
H_0: The proportion of Liberal supporters is greater than or equal to 0.40
H_a: The proportion of Liberal supporters is less than 0.40.
H_0: $\pi \geq 0.40$
H_a: $\pi < 0.40$

Chapter Review

In a **hypothesis test**, sample data is evaluated in order to arrive at a decision about some type of claim about a population parameter, such as the mean or proportion. If the sample provides strong evidence to the contrary of the original claim, then the claim can be rejected in favour of the new claim. In a hypothesis test, we:

1. Evaluate the **null hypothesis**, typically denoted with *H0*. The null is not rejected unless the hypothesis test shows otherwise. The null statement must always contain some form of equality $(= , \leq \text{ or } \geq)$
2. Always write the **alternative hypothesis**, typically denoted with *Ha* or *H1*, using not equal, less than or greater than symbols, i.e., $(\neq , < , \text{or } >)$.
3. If we reject the null hypothesis, then we can assume there is enough evidence to support the alternative hypothesis.
4. Never state that a claim under the null hypothesis is proven true or false. Keep in mind the underlying fact that hypothesis testing is based on probability laws; therefore, we can talk only in terms of non-absolute certainties.

You are testing that the mean speed of your cable Internet connection is more than three Megabits per second. What is the random variable? Describe in words.

The random variable is the mean Internet speed in Megabits per second.

Canadian families have an average of two children. What is the random variable? Describe in words.

The random variable is the mean number of children a Canadian family has.

A sociologist claims the probability that a person picked at random visting the CN Tower in Toronto is a tourist is 0.83. You want to test to see if the proportion is actually less. What is the random variable? Describe in words.

The random variable is the proportion of people who are tourists picked at random at the CN Tower.

In a population of fish, approximately 42% are female. A test is conducted to see if, in fact, the proportion is less. State the null and alternative hypotheses.

---

1. *Ho*: $\pi = 0.42$
2. *Ha*: $\pi < 0.42$

Homework

Some of the following statements refer to the null hypothesis, some to the alternate hypothesis. Hint: pay attention to whether the statement states or implies an equality. If so, it refers to the null hypothesis.

State the null hypothesis, H_0, and the alternative hypothesis. H_a, in terms of the appropriate parameter (μ or π).

1. The mean number of years Canadians work before retiring is 34.
2. At most 60% of Canadians vote in federal elections.
3. The mean starting salary for U of A graduates is at least $100,000 per year.
4. Twenty-nine percent of high school seniors get drunk each month.
5. Fewer than 5% of adults ride the bus to work in Calgary.
6. The mean number of cars a person owns in her lifetime is not more than ten.
7. About half of Canadians prefer to live away from cities, given the choice.
8. Europeans have a mean paid vacation each year of six weeks.
9. The chance of developing breast cancer is under 11% for women.

10. Private universities' mean tuition cost is more than $20,000 per year.

---

1. H_0: $\mu = 34$; H_a: $\mu \neq 34$
2. H_0: $\pi \leq 0.60$; H_a: $\pi > 0.60$
3. H_0: $\mu \geq 100{,}000$; H_a: $\mu < 100{,}000$
4. H_0: $\pi = 0.29$; H_a: $\pi \neq 0.29$
5. H_0: $\pi = 0.05$; H_a: $\pi < 0.05$
6. H_0: $\mu \leq 10$; H_a: $\mu > 10$
7. H_0: $\pi = 0.50$; H_a: $\pi \neq 0.50$
8. H_0: $\mu = 6$; H_a: $\mu \neq 6$
9. H_0: $\pi \geq 0.11$; H_a: $\pi < 0.11$
10. H_0: $\mu \leq 20{,}000$; H_a: $\mu > 20{,}000$

A statistics instructor believes that fewer than 20% of Lethbridge Community College (LCC) students attended the opening night midnight showing of the latest Harry Potter movie. She surveys 84 of her students and finds that 11 attended the midnight showing. An appropriate alternative hypothesis is:

1. $\pi = 0.20$
2. $\pi > 0.20$
3. $\pi < 0.20$
4. $\pi \leq 0.20$

---

c

References

Data from the National Institute of Mental Health. Available online at http://www.nimh.nih.gov/publicat/depression.cfm.

Glossary

Hypothesis

a statement about the value of a population parameter, in case of two hypotheses, the statement assumed to be true is called the null hypothesis (notation H_0) and the contradictory statement is called the alternative hypothesis (notation H_a).

Errors and Choosing a Level of Significance

Errors in Hypothesis Testing

Any time we reject a claim (Ho), there is a possibility we were wrong. Rejecting an Ho that is actually true is known as a Type I Error. For example, when we send someone who is innocent to jail, we have committed a Type I error; we have rejected a null hypothesis that is actually true. If making such an error is costly (financially, to someone's well being or otherwise), we would want to severely limit the possibility of this kind of error from occurring. Conversely, any time we fail to reject a claim (Ho), there is also possibility we were wrong. If a claim is actually false but we fail to reject that claim, we have committed what is known as a Type II Error. If a Type II error is deemed to be more costly than a Type I error we would strive to limit the possibility of this kind of error from occurring.

How? Recall from Chapter 7 that we can decide in advance how confident we wish to be in our confidence interval estimates. We do something similar in hypothesis testing by choosing what is known as a level of significance. The level of significance, identified by the Greek letter alpha α, is simply 1 minus our level of confidence. So a 95% level of confidence has a corresponding level of significance of 5%. In terms of a Type I error, an alpha of 5% is the probability that our test could lead to rejecting a null hypothesis that is actually true. As mentioned above, if a Type I error is deemed very costly, we may wish to reduce alpha to as low as 1%. This means that the probability our test could lead to a rejecting of a null hypothesis that is actually true is only 1%. So why not set alpha at 0%? That way we would never make a Type I error. Setting alpha at o% would require us to have perfect evidence before we would be able to reject the null hypothesis. Imagine if this were the case in a court trial. The judge would instruct the jury not to convict unless the evidence of guilt was absolutely perfect and all jury members were 100% certain of the defendant's guilt. If this were the case, we would rarely send anyone to jail and we would have a lot more dangerous people roaming our streets. In short, it is unreasonable to demand that sample evidence provide perfect proof against the null hypothesis.

A Type II error is known by the Greek letter beta β. Unfortunately, we cannot predetermine beta in the same way we do with alpha, but we do know the two types of errors share an inverse relationship: the lower we set alpha, the higher beta becomes and vice versa. Back to our courtroom example. If we reduced to probability of making a Type I error to 0, as we said, we would allow almost everyone to go free, even if they were guilty, for lack of perfect evidence. When we send a person guilty of a serious crime back on the street, we have committed a Type

II error--we have failed to reject a null hypothesis that is actually false. And since the judge set alpha at 0 (that is, he demanded perfect proof of guilt before being willing to convict), he has sent beta soaring. Almost no one will be convicted. Since we can't set beta in advance, we must set our level of alpha high (for example, 10%) to minimize a Type II error.

To illustrate further, let's say a certain medical condition is easy to treat with a drug that poses little danger and has few side effects. Let's also say this condition is relatively hard to diagnose because it shares symptoms with several other conditions. A stomach ulcer is one example. The doctor tests you for an ulcer by looking for evidence, such as pressing on your stomach and discussing your symptoms. As best as she can tell, she decides there is a good chance you have an ulcer. She prescribes a drug and off you go. After one month, your symptoms persist and so you re-visit the doctor who then rules out her earlier diagnosis in favour of a new one. What has happened here is that in her initial diagnosis the doctor had made a Type I error. She has rejected the null hypothesis (that you don't have an ulcer) in favour of the alternative hypothesis that you do have an ulcer. As it turns out, she was wrong. She prescribed a drug that would not help you for a condition you do not have. Before getting too anxious about the medical system, keep in mind that this is a fairly common practice in diagnosing relatively benign conditions that can be treated easily. The old saying, "Take two aspirin and call me in the morning" sums this approach up well. Recall that the doctor diagnosed your ulcer by taking in only a few pieces of evidence: talking to you and pressing on your stomach. In other words, she was willing to reject the null hypothesis on relatively weak evidence. Why? Because she knew that the prescription might help, and even if it didn't it would do you little harm. And since it didn't help you after a month, she can now rule out an ulcer and focus on other, possibly less benign, conditions. Keep in mind that if she had set alpha low, she likely would not have misdiagnosed you, but she would also have sought much stronger evidence--possibly even invasive exploratory surgery--before being willing to reject the null hypothesis. Obviously, in this case it made much more sense to risk a Type II error and treat you for a condition that you don't actually have.

Summary

When you perform a hypothesis test, there are actually four possible outcomes depending on the actual truth (or falseness) of the null hypothesis H_0 and the decision to reject or not. The outcomes are summarized in the following table:

STATISTICAL DECISION	***H0* IS ACTUALLY...**	
	True	False
Cannot reject *H0*	Correct Outcome	Type II error
Cannot accept *H0*	Type I Error	Correct Outcome

The four possible outcomes in the table are:

1. The decision is **cannot reject *H0*** when ***H0* is true (correct decision).**
2. The decision is **cannot accept *H0*** when ***H0* is true** (incorrect decision known as a**Type I error**). This case is described as "rejecting a good null". As we will see later, it is this type of error that we will guard against by setting the probability of making such an error. The goal is to NOT take an action that is an error.
3. The decision is **cannot reject *H0*** when, in fact, ***H0* is false** (incorrect decision known as a **Type II error**). This is called "accepting a false null". In this situation you have allowed the status quo to remain in force when it should be overturned. As we will see, the null hypothesis has the advantage in competition with the alternative.
4. The decision is **cannot accept *H0*** when ***H0* is false** (**correct decision** whose probability is called the **Power of the Test**).

Each of the errors occurs with a particular probability. The Greek letters α and β represent the probabilities.

α = probability of a Type I error = ***P*(Type I error)** = probability of rejecting the null hypothesis when the null hypothesis is true.

β = probability of a Type II error = ***P*(Type II error)** = probability of not rejecting the null hypothesis when the null hypothesis is false.

The following are examples of Type I and Type II errors.

Suppose the null hypothesis, *Ho*, is: Frank's rock climbing equipment is safe.
Type I error: Frank thinks that his rock climbing equipment may not be safe when, in fact, it really is safe. **Type II error**: Frank thinks that his rock climbing equipment may be safe when, in fact, it is not safe.
$\boldsymbol{\alpha}$ = **probability** that Frank thinks his rock climbing equipment may not be safe when, in fact, it really is safe. $\boldsymbol{\beta}$ = **probability** that Frank thinks his rock climbing equipment may be safe when, in fact, it is not safe.
Notice that, in this case, the error with the greater consequence is the Type II error. (If Frank thinks

his rock climbing equipment is safe, he will go ahead and use it.)
This is a situation described as "accepting a false null".

Try It

Suppose the null hypothesis, H_0, is: the blood cultures contain no traces of pathogen X. State the Type I and Type II errors.

Type I error: The researcher thinks the blood cultures do contain traces of pathogen X, when in fact, they do not.

Type II error: The researcher thinks the blood cultures do not contain traces of pathogen X, when in fact, they do.

Try It

Suppose the null hypothesis, H_0, is: a patient is not sick. Which type of error has the greater consequence, Type I or Type II?

The error with the greater consequence is the Type II error: the patient will be thought well when, in fact, he is sick, so he will not get treatment.

Try It

“Red tide” is a bloom of poison-producing algae–a few different species of a class of plankton called dinoflagellates. When the weather and water conditions cause these blooms, shellfish such as clams living in the area develop dangerous levels of a paralysis-inducing toxin. In Massachusetts, the Division of Marine Fisheries (DMF) monitors levels of the toxin in shellfish by regular sampling of shellfish along the coastline. If the mean level of toxin in clams exceeds 800 μg (micrograms) of toxin per kg of clam meat in any area, clam harvesting is banned there until the bloom is over and levels of toxin in clams subside. Describe both a Type I and a Type II error in this context, and state which error has the greater consequence.

In this scenario, an appropriate null hypothesis would be*Ho*: the mean level of toxins is at most 800 μg, $H_0 : \mu_0 \leq 800\ \mu$g.

Type I error: The DMF believes that toxin levels are still too high when, in fact, toxin levels are at most 800 μg. The DMF continues the harvesting ban.

Type II error: The DMF believes that toxin levels are within acceptable levels (are at least 800 μg) when, in fact, toxin levels are still too high (more than 800 μg). The DMF lifts the harvesting ban. This error could be the most serious. If the ban is lifted and clams are still toxic, consumers could possibly eat tainted food.

In summary, the more dangerous error would be to commit a Type II error, because this error involves the availability of tainted clams for consumption.

Try It

Determine both Type I and Type II errors for the following scenario:
Assume a null hypothesis, H_0, that states the percentage of adults with jobs is at least 88%.

Identify the Type I and Type II errors from these four statements.

1. Not to reject the null hypothesis that the percentage of adults who have jobs is at least 88% when that percentage is actually less than 88%
2. Not to reject the null hypothesis that the percentage of adults who have jobs is at least 88% when the percentage is actually at least 88%.
3. Reject the null hypothesis that the percentage of adults who have jobs is at least 88% when the percentage is actually at least 88%.
4. Reject the null hypothesis that the percentage of adults who have jobs is at least 88% when that percentage is actually less than 88%.

Type I error: c

Type I error: b

Chapter Review

In every hypothesis test, the outcomes are dependent on a correct interpretation of the data. Incorrect calculations or misunderstood summary statistics can yield errors that affect the results. A **Type I** error occurs when a true null hypothesis is rejected. A **Type II error** occurs when a false null hypothesis is not rejected.

The probabilities of these errors are denoted by the Greek letters α and β, for a Type I and a Type II error respectively.

The mean price of mid-sized cars in a region is $32,000. A test is conducted to see if the claim is true. State the Type I and Type II errors in complete sentences.

Type I: The mean price of mid-sized cars is $32,000, but we conclude that it is not $32,000.

Type II: The mean price of mid-sized cars is not $32,000, but we conclude that it is $32,000.

For Exercise 9.12, what are α and β in words?

$\alpha =$ the probability that you think the bag cannot withstand -15 degrees F, when in fact it can

$\beta =$ the probability that you think the bag can withstand -15 degrees F, when in fact it cannot

A group of doctors is deciding whether or not to perform an operation. Suppose the null hypothesis, *Ho*, is: the surgical procedure will go well. State the Type I and Type II errors in complete sentences.

Type I: The procedure will go well, but the doctors think it will not.

Type II: The procedure will not go well, but the doctors think it will.

Homework

State the Type I and Type II errors in complete sentences given the following statements.

1. The mean number of years Americans work before retiring is 34.
2. At most 60% of Americans vote in

presidential elections.

3. The mean starting salary for San Jose State University graduates is at least $100,000 per year.
4. Twenty-nine percent of high school seniors get drunk each month.
5. Fewer than 5% of adults ride the bus to work in Los Angeles.
6. The mean number of cars a person owns in his or her lifetime is not more than ten.
7. About half of Americans prefer to live away from cities, given the choice.
8. Europeans have a mean paid vacation each year of six weeks.
9. The chance of developing breast cancer is under 11% for women.
10. Private universities mean tuition cost is more than $20,000 per year.

---

1. Type I error: We conclude that the mean is not 34 years, when it really is 34 years. Type II error: We conclude that the mean is 34 years, when in fact it really is not 34 years.
2. Type I error: We conclude that more than 60% of Americans vote in presidential elections, when the actual percentage is at most 60%.Type II error: We conclude that at most 60% of Americans vote in presidential elections when, in fact, more

than 60% do.

3. Type I error: We conclude that the mean starting salary is less than $100,000, when it really is at least $100,000. Type II error: We conclude that the mean starting salary is at least $100,000 when, in fact, it is less than $100,000.
4. Type I error: We conclude that the proportion of high school seniors who get drunk each month is not 29%, when it really is 29%. Type II error: We conclude that the proportion of high school seniors who get drunk each month is 29% when, in fact, it is not 29%.
5. Type I error: We conclude that fewer than 5% of adults ride the bus to work in Los Angeles, when the percentage that do is really 5% or more. Type II error: We conclude that 5% or more adults ride the bus to work in Los Angeles when, in fact, fewer that 5% do.
6. Type I error: We conclude that the mean number of cars a person owns in his or her lifetime is more than 10, when in reality it is not more than 10. Type II error: We conclude that the mean number of cars a person owns in his or her lifetime is not more than 10 when, in fact, it is more than 10.
7. Type I error: We conclude that the proportion of Americans who prefer to live

away from cities is not about half, though the actual proportion is about half. Type II error: We conclude that the proportion of Americans who prefer to live away from cities is half when, in fact, it is not half.

8. Type I error: We conclude that the duration of paid vacations each year for Europeans is not six weeks, when in fact it is six weeks. Type II error: We conclude that the duration of paid vacations each year for Europeans is six weeks when, in fact, it is not.
9. Type I error: We conclude that the proportion is less than 11%, when it is really at least 11%. Type II error: We conclude that the proportion of women who develop breast cancer is at least 11%, when in fact it is less than 11%.
10. Type I error: We conclude that the average tuition cost at private universities is more than $20,000, though in reality it is at most $20,000. Type II error: We conclude that the average tuition cost at private universities is at most $20,000 when, in fact, it is more than $20,000.

For statements a-j in Exercise 9.109, answer the following in complete sentences.

1. State a consequence of committing a Type

I error.
2. State a consequence of committing a Type II error.

When a new drug is created, the pharmaceutical company must subject it to testing before receiving the necessary permission from the Food and Drug Administration (FDA) to market the drug. Suppose the null hypothesis is “the drug is unsafe.” What is the Type II Error?

1. To conclude the drug is safe when in, fact, it is unsafe.
2. Not to conclude the drug is safe when, in fact, it is safe.
3. To conclude the drug is safe when, in fact, it is safe.
4. Not to conclude the drug is unsafe when, in fact, it is unsafe.

---

b

It is believed that Lake Tahoe Community College (LTCC) Intermediate Algebra students get less than seven hours of sleep per night, on average. A survey of 22 LTCC Intermediate Algebra students generated a mean of 7.24

hours with a standard deviation of 1.93 hours. At a level of significance of 5%, do LTCC Intermediate Algebra students get less than seven hours of sleep per night, on average?

The Type II error is not to reject that the mean number of hours of sleep LTCC students get per night is at least seven when, in fact, the mean number of hours

1. is more than seven hours.
2. is at most seven hours.
3. is at least seven hours.
4. is less than seven hours.

---

d

Glossary

Type I Error
: The decision is to reject the null hypothesis when, in fact, the null hypothesis is true.

Type II Error
: The decision is not to reject the null hypothesis when, in fact, the null hypothesis is false.

The Eight-Step Hypothesis Test
This module covers the formal hypothesis test using an eight-step approach with an emphasis on p-values.

P-values and the Level of Significance

Once you have set out your null and alternative hypothesis, you need to determine how strong your sample evidence must be before you would be willing to reject the null hypothesis in favour of the alternative hypothesis. The required strength of evidence is defined by the level of significance (α).

Once your level of significance has been set, you can then examine your sample evidence to determine its strength, as measured by its p-value This process will be discussed below.

Ethical Implications of Choosing a Level of Significance

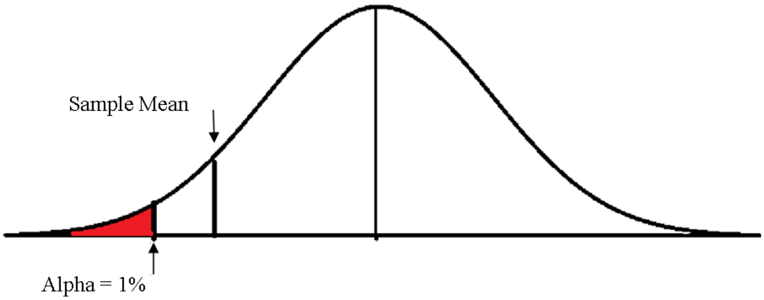
Once you have set out your null and alternative hypothesis, you need to determine how strong your sample evidence must be before you would be confident in rejecting the null hypothesis in favour of the alternative hypothesis. The required strength of evidence is defined by the level of significance (α).

Typically values for alpha range from 1% to 10% and will vary depending on a number of factors, including conventions set by a particular industry or discipline and the relative risks of a Type I versus a Type II error, as discussed in the previous section. In many cases, the choice of alpha may be left up to the analyst. Unfortunately, without a peer review process, some analysts may be tempted to set alpha in a way that will support his or her desired conclusion.

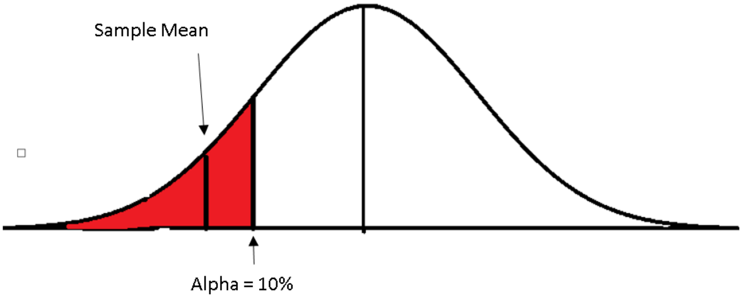
For example, if a pharmaceutical company stands to make millions of dollars on a new drug, it obviously has a vested interest in offering *proof* that the drug is effective. The null hypothesis is that the drug is not effective; and the aternative is that it is. But what if the proof, as discovered by several rounds of double-blind tests, turns out to be rather weak? This would normally lead the researcher to decide not to reject the null hypothesis and conclude that the sample evidence is insufficiently strong for the drug to be considered a success. If this were the conclusion, the drug should not be approved as an effective treatment. But a company with millions already invested in the drug may be strongly determined to see it to market, in spite of the test results. An unethical approach might be to simply move the *goal posts* to make it easier to reject the null hypothesis (i.e. to make the proof look stronger than it is).

These *goal posts*, of course, are defined by the level of significance. In much scientific testing, the level of significance is typically set at 1%, which means the sample evidence must be very strong before a null hypothesis can be rejected. In this case, moving the goal posts could mean setting the level of significance as high as 10%. This higher level of significance, as we shall see below, allows for weaker evidence to be used in support of an alternative hypothesis.

In Figure 1 below, alpha has been set at 1%. As you can see, the sample evidence fails to cross over the *goal posts* set by alpha and we would thus reach a fail to reject of the null hypothesis. The sample evidence is not strong enough.



In the following figure we have moved the *goal posts* by setting alpha at 10%, making it easier to reject the null hypothesis. As you can see, the sample evidence now is strong enough to lead us to reject the null hypothesis. Of course, in truth the evidence has not changed, but in the first instance we fail to reject the null and in the second we do reject the null.



Thankfully, at least when it comes to pharmaceutical testing, there are objective, government regulated standards that cannot be easily manipulated by vested interests. However, there are instances where the researcher is in control of choosing the level of significance. When this is the case, the choice should be made ethically and with an honest consideration of the implications of Type I and Type II errors.

As a final note, the level of significance should never be chosen *after* the sample evidence has been measured. This would be akin to allowing the home team to determine where the goal posts are after the game has already begun.

Examining the Sample Evidence

Once the level of significance has been set, you can look more closely at the sample evidence to determine how strong it is. As discussed earlier, this evidence is first measured by determining how far away your sample mean or proportion is from the hypothesized mean or proportion. The measuring stick we use is called a Z-score or a t-score, which is simply the number of standard deviations our sample mean or proportion lies from the hypothesized middle of the sampling distribution.

Recall from earlier in this chapter the example we looked at regarding sleep habits. We hypothesized that the mean number of hours adults sleep per night is 8. We then gathered sample evidence, where the sample mean was 7.5 and the standard deviation was 1.4 hours. The sampling distribution for this scenario would then have a hypothesized middle of 8 and a standard error of 0.20 (i.e. 1.4/sqrt50)

Does a sample mean of 7.5 provide sufficient proof that the true population mean is less than 8? To investigate, we must first determine our level of significance. For now, we will use the default of 5%. This means that if our sample mean falls into the lower 5% of the tail, it will be considered strong evidence against the null hypothesis. We can now measure how many standard deviations (Z-scores, since we are working with a large sample) 7.5 is from 8. This measurement is often called the *test statistic*. You may see it written as Z* or t*.

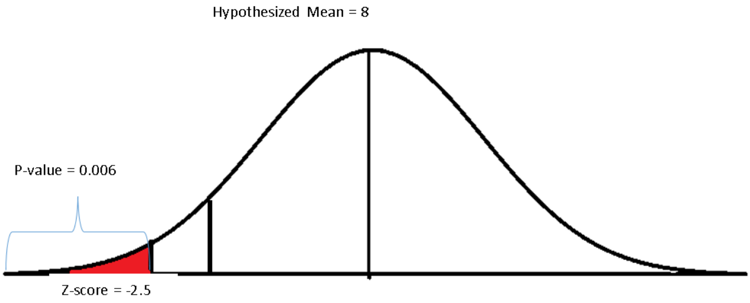
Using our standardizing formula, we get Z* = 7.5 – 8.0 / (1.4/√50). The resulting Z-score is -2.5

(rounded to one decimal). Based on the empirical rule we know that any value with a Z-score of 2.5 (as an absolute value) would fall well out into the lower or upper 5% of the tail and would thus be considered an *unlikely* observation. That is, very few sample means taken from a population with a mean of 8 would have such a high Z-score.

Our decision, in this case would be to reject the null hypothesis (that the mean number of hours adults sleep is 8) in favour of the alternative hypothesis (that the mean number of hours adults sleep is less than 8). Keep in mind we have not proven they only sleep 7.5; this is never what we sought to prove. We only sought to prove that they sleep *less* than 8 hours. Our sample mean of 7.5 is our evidence against the null hypothesis. As it turned out, the empirical rule helped us conclude that a sample mean of 7.5 would be a very unlikely finding if the true population mean were actually 8, which is why we rejected the null hypothesis.

Measuring Sample Evidence with P-Values

While using Z-scores and t-scores can lead us to a correct decision, a more common and precise measuring tool is preferred, called a p-value. To find the p-value of a sample mean or proportion we simply need to convert the test statistic into a probability. Specifically, the p-value seen below is the probability of getting sample mean of 7.5 or less from a population whose true mean is 8. As you can see, the resulting p-value is extremely small, meaning that such an outcome would be extremely unlikely (well under a probability of 5%) to occur if the true mean is 8.



Be careful! The p-value is not the probability that the null hypothesis is true. It is the probability that our sample mean could have come from a population in which the null hypothesis is true. And since this probability is so small, we must conclude the null hypothesis in not true. In other words, our sample mean is what is considered an *unlikely event*.

P-values and Unlikely Events

As a final example, suppose Didi and Ali are at a birthday party of a very wealthy friend. They hurry to be first in line to grab a prize from a tall basket that they cannot see inside because they will be blindfolded. There are 200 plastic bubbles in the basket and Didi and Ali have been told that there is only one with a $100 bill. Didi is the first person to reach into the basket and pull out a bubble. Her bubble contains a $100 bill. The probability of this happening is $1/200 = 0.005$.

In statistical language, 0.005 is akin to a p-value. Because this occurrence was unlikely to have happened if there truly is only one $100 bill in the basket, Ali can conclude that what the two of them were told was wrong and there are actually more $100 bills in the basket. A "rare event" has occurred (Didi getting the $100 bill), so Ali doubts the assumption about only one $100 bill being in the basket.

The Decision and Conclusion

Once you have determined the p-value associated with a sample mean or proportion, the next step is to compare that p-value to the original level of signficnce..

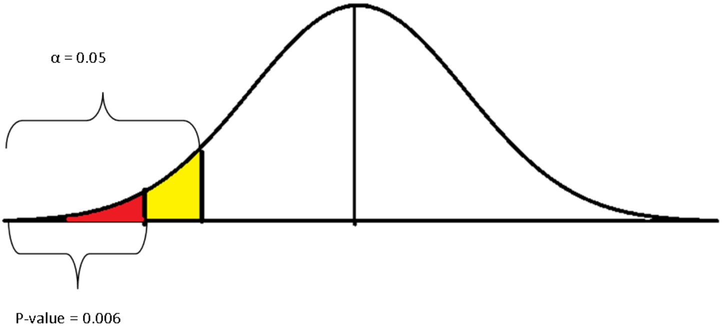
When you make a **decision** to reject or not reject *H0*, do as follows:

If p-value $< \alpha$, reject *H0*. The evidence provided by the sample data is significant. There is sufficient evidence to conclude that *H0* is an incorrect belief and that the **alternative hypothesis**, *Ha*, may be correct.

If p-value $\alpha \geq$, do not reject *H0*. The evidence provided by the sample data is not significant.There is not sufficient evidence to conclude that the alternative hypothesis,*Ha*, may be correct.

When you "do not reject *H0*", it does not mean that you have proven that *H0* is true. It simply means that the sample data have **failed** to provide sufficient evidence to cast serious doubt about the truthfulness of *Ho*.

The figure below illustrates a P-value of 0.006 and a chosen level of significance of 0.05. As you can see, the p-value is much smaller than alpha (further out into the tail), which indicates strong evidence against the null hypothesis.



Conclusion: After you make your decision, write a thoughtful **conclusion** about the hypotheses in terms of the given problem, making specific reference to the context. The example below should serve as a summary and a guide for conducting a full eight-step hypothesis test on a population mean or proportion.

Conducting the Hypothesis Test

In this course, we stress an eight-step process for conducting a hypothesis test.

1. **Determine and record Ho and Ha**, as discussed earlier in this chapter.
2. **Record the sample evidence** that you will be using to challenge Ho. For a means test, your evidence will consist of the sample mean, the sample (or population) standard deviation, and the sample size. For a proportions test, your evidence will consist of the sample proportion and the sample size.
3. **State the test considerations**. Looking at the sample evidence and any stated assumptions, determine the correct test procedure.
4. **State the required strength of evidence.** Consider the implications of a Type I vs. a Type II error in choosing your level of significance, as well as any ethical considerations.
5. **Calculate the test statistic**. Using the sample evidence, compute Z* or t* and the associated p-value.
6. **Discuss what the p-value measures in context** and whether the test statistic can be considered an unlikely or a likely event within the context of the problem.
7. **Make a decision.** Compare the test statistic

(the p-value) to the required strength of evidence (alpha) and determine if you can reject or fail to reject the null hypothesis.
8. **Offer a concluding sentence.** Using accessible language summarize your conclusion in sentence form within the context of the problem.

Example 1

Suppose Irene, who owns a top bakery in the city, claims that she has the best bread in the city by any measure. Not only is her bread the tastiest, it is also the fluffiest and the tallest, averaging 15 cm in height. Another baker, Jose, wishes to challenge Irene's claim that her bread is the tallest. As evidence he will provide a sample of 40 randomly selected loaves of bread and have their heights measured in his attempt to prove that his bread heights actually exceed 15 cm, on average. In doing so, he obtains a sample mean bread height of 15.5 cm. He also knows from baking thousands of loaves that his variation is very low: specifically the standard deviation is 0.9 cm.

Step One

The null and alternative hypotheses are as follows:

Ho: $\mu = 15$

Ha: $\mu > 15$

Step Two

The sample evidence is as follows: sample mean = 15.5; population standard deviation = 0.9; sample size = 40.

Step Three

The test considerations are as follows: We are using a large sample ($n > 30$) to conduct a means test. This will require a sampling distribution of the mean, which central limit theorem says will be approximately normally shaped since our sample size exceeds 30. We will therefore do a Z-based test.

Step Four

The required strength of evidence can now be determined by considering the implications of a Type I vs. a Type II error. In this context, Jose will make a Type I error if he concludes that his bread heights average more than 15 cm when in fact they do not. He will make a Type II error if he concludes that his bread heights do not average more than 15 cm when in fact they do. Which error is worse will depend on where you are standing. Jose would consider a Type II error worse, whilst Irene would consider a Type I error worse. To be fair, we will choose a level of significance of 5%, which is generally considered a good balance between the two types of errors.

Reject Ho if p-value < 0.05

Step Five

Compute the test statistics as follows:

Z* = 15.5-15/(0.9/√40) = 3.51

p-value = 0.0002

NOTE: The Excel function for computing a p-value is as follows:

=1-NORM.S.DIST(3.51,1)

Step Six

Interpret the p-value in the context of the problem. Under the assumption that Jose's bread is no taller than Irene's (this his bread averages only 15 cm), the probability of obtaining a sample of 40 with a mean of 15.5 cm (or more) is only 0.0002 or 0.02%, which makes it a very unlikely event.

Step Seven

Make a decision by comparing your p-value to your level of significance.

Since the p-value (0.0002) < 0.05, we can reject the null hypothesis.

Step Eight

Offer a final conclusion in sentence form: Therefore we can conclude that Jose's bread averages more than 15 cm and is indeed taller than Irene's.

Practice Question One

> An auditing firm is looking at the travel expense claims for a large book retailer. The retailer's books suggest that their average (μ) travel expenses was $1200 per person per year. A sample of 64 random expense claims revealed average of $1300. The population σ, based on an earlier comprehensive audit, is $400. The sample suggests the books have under-exaggerated the expense claims. Identify the Null and Alternative Hypotheses.

> Ho: $\mu = \$1200$; Ha: $\mu > \$1200$
>
> State your evidence.

> A sample of 64 random expense claims revealed average of $1300. The population σ, based on an earlier comprehensive audit, is $400.
>
> Identify all test considerations and then determine the appropriate test.

We are investigating a hypothesis about a population mean using a large sample. Central Limit Theorem says the sampling distribution will be normally shaped for sample sizes over 30. Thus we will conduct a Z-based test.

Consider the implications of both Type I and Type II errors and then decide on an appropriate level of significance. State your decision rule.

A Type I error in this case would be for the auditor to accuse the bookstore of exaggerating its expense claims when in fact it has not. A Type II error in this case would be for the auditor to not accuse the bookstore of exaggerating its expense claims when in fact it has.A Type I error could lead to a wrongful conviction for tax fraud so it would be best to minimize the likelihood of making this type of error. Alpha should be set at 1% (or at most 5%). Reject Ho if P-value < 0.01.

Calculate the test statistics.

$Z^* = 1300-1200/(400/\sqrt{64}) = 2.00$; P-value = 0.0228

Define what the p-value is measuring in the context of the problem.

Our P-value is 0.0228. This means that the probability of getting a sample mean of $1300 (or more) from a population with a mean of $1200 is 2.28%. Given our level of significance, this would be considered a not unlikely event. The P-value $>$ 0.01, so we will fail to reject the null hypothesis.

Make a decision.

Our P-value of 0.0228 is less than alpha of 0.01, we can reject the null hypothesis.

Draw a final conclusion in sentence form.

There is insufficient evidence to indicate that the average yearly travel expenditures per person per year is greater than $1200.

Practice Question Two

A charitable organization wanted to see if a new form of mail marketing would change the percentage of people who replied. In the past the percentage of people who would reply to mail marketing was 1 in 175. A sample of 2000 letters was sent out. A total of 20 people responded. Is there any significant change in the percentage of respondents? Identify the null and alternative hypotheses.

HO: $\pi = 0.0057$ HA: $\pi \neq 0.0057$

State your evidence.

$n = 2000$; number of success $= 20$

Identify all test considerations and then determine the appropriate test.

We are testing if there has been a change in the proportion of successes within a population. To ensure that a large sample z-test is valid, we must ensure that both np and n(1-p) > 5. In this case $np = = 6$ and $n(1\text{-}p) = 1980$, so central limit theorem says the sample distribution of sample proportions will follow and approximately normal shape. Thus we will conduct a z-based test.

Consider the implications of both Type I and Type II errors and then decide on an appropriate level of significance. State your decision rule.

A Type I error in this context would conclude the campaign has been successful when in fact it has not. A Type II error would conclude the campaign has not been successful when in fact it has. If the campaign is costly, it would be better to err on the side of making a Type II error over a Type I error. Therefore we will set alpha at 10%. Reject Ho if the p-value < 0.10.

Calculate the test statistics.

Z* = (0.01 – 0.0057)/ sqrt((0.0057*0.9943)/2000) = 2.54; P-value = 0.011

Define what the p-value is measuring in the context of the problem.

Our P-value is 0.011. This means that the probability of getting a sample proportion of 0.01 (or more) from a population with a proportion of 0.0057 is only 1.1%. Given our level of significance, this would be considered an unlikely event

Make a decision by comparing the P-value to α.

Since the p-value of 0.011 is less than alpha of 0.10, we will reject the null hypothesis.

Draw a final conclusion in sentence form.

There is sufficient evidence to indicate that the proportion of responses differs as a result of the marketing campaign.

Practice Question Three

Charter Air claims that its new executive boarding service has improved the time it takes for business passengers to purchase tickets, store luggage and board the plane. They believe that is less than the previous time of 12 minutes. A sample of 9 customers of this new exclusive service indicates the that the mean is 9.3 minutes with a standard deviation of 3.32 minutes. Previous studies have revealed that boarding times tend to follow a normal distribution. Identify the null and alternative hypotheses.

Ho: $\mu = 12$; Ha: $\mu < 12$

State your evidence.

A sample of 9 randomly chosen customers' boarding times reveals: $n = 9$; mean $= 9.3$; sample standard deviation $= 3.32$.

Identify all test considerations and then determine the appropriate test.

We are investigating a hypothesis about a population mean using a small sample. Central Limit Theorem does not apply to small samples, but we can expect the sampling distribution to be normally shaped if the population is also normal. This has been confirmed through previous studies. Thus we will conduct a t-based test.

Consider the implications of both Type I and Type II errors and then decide on an appropriate level of significance. State your decision rule.

A Type I error in this case would be for Charter

Air to claim their boarding time is less than 12 minutes, when in fact it is not. A Type II error in this case would be for Charter Air not to claim their boarding time is less than 12 minutes, when in fact it is. A Type I error could lead to false advertising, which has both ethical and legal implications, so it would be best to minimize the likelihood of making this type of error. Alpha should be set at 1% (or at most 5%). Reject Ho if P-value < 0.01.

Calculate the test statistics.

---

t* = 9.3-12/(3.32/√9) = -2.43; P-value = 0.0203

Make a decision by comparing the P-value to α.

The P-value > 0.01, so we will fail to reject the null hypothesis.

Define what the p-value is measuring in the context of the problem.

Our P-value is 0.0203. This means that the probability of getting a sample mean of 9.3

minutes (or less) from a population with a mean of 12 minutes is 2.03%. Given our level of significance, this would be considered a not unlikely event.

Make a decision by comparing the p-value to alpha

The P-value of $0.0203 > 0.01$, so we will fail to reject the null hypothesis.

Draw a final conclusion in sentence form.

There is insufficient evidence to indicate that the mean boarding time is less than 12 minutes.

Practice Questions for Confidence Intervals and Hypothesis Tests
Eight practice questions for the end of unit on one-sample hypothesis tests and confidence intervals.

Practice Questions for Confidence Intervals and Hypothesis Tests

These questions were derived from Lyryx Learning, Business Statistics I -- MGMT 2262 -- Mt Royal University -- Version 2016 Revision A. OpenStax CNX. Sep 8, 2016 http://cnx.org/contents/f3aefa9e-58d2-41ea-969f-04dc2cb04c82@5.5.

If a question has a set of data, please see the course site for the Excel file.

Solutions are at the end of the chapter.

1. Question 1: The Specific Absorption Rate (SAR) for a cell phone measures the amount of radio frequency (RF) energy absorbed by the user's

body when using the handset. Every cell phone emits RF energy. Different phone models have different SAR measures. To receive certification from the Federal Communications Commission (FCC) for sale in the United States, the SAR level for a cell phone must be no more than 1.6 watts per kilogram. Table 7.1 shows the highest SAR level for a random selection of cell phone models as measured by the FCC. A recent study has shown that if a cell phone's SAR level exceeds 0.9 watts per kilogram, there is an increased chance of brain tumours for those that use this phone[footnote] An advocacy group wants to use this new study to petition the FCC to change their regulations around the current allowable SAR levels.

Phone model	SAR	Phone model	SAR	Phone model	SAR
Apple iPhone 4S	1.11	LG Ally	1.36	Pantech Laser	0.74
BlackBerry Pearl	1.48	LG AX275	1.34	Samsung Character	0.5
BlackBerry Tour	1.43	LG Cosmos	1.18	Samsung Epic 4G Touch	0.4
Cricket TXTM8	1.3	LG CU515	1.3	Samsung M240	0.867
HP/Palm Centro	1.09	LG Trax CU575	1.26	Samsung Messenger III	0.68

HTC One V	0.455	Motorola Q9h	1.29	Samsung Nexus S	0.51
HTC Touch Pro 2	1.41	Motorola Razr2 V8	0.36	Samsung SGH-A227	1.13
Huawei M835 Ideos	0.82	Motorola Razr2 V9	0.52	SGH-a107 GoPhone	0.3
Kyocera DuraPlus	0.78	Motorola V195s	1.6	Sony W350a	1.48

1. What is the variable being studied? Categorize it. Based on this, what descriptive statistic (mean or proportion) is best for this situation?
2. Is it appropriate to assume that the sampling distribution is normal? Explain your reasoning and provide evidence for your choice. Regardless of your answer in b), assume that the sampling distribution is normal for the remaining questions.
3. The advocacy group will go forward with their petition if they can show that, on average, cell phones have SAR rates that exceed 0.9 watts per kg. This advocacy group is run by an administrator who is very risk averse (meaning they will only go forward with the petition if there is a lot of evidence). Determine whether the advocacy group should go forward with their petition by performing an

appropriate eight-step hypothesis test.

4. Find a confidence interval for the true (population) mean of the Specific Absorption Rates (SARs) for cell phones. Choose a confidence level that complements the level of significance you have chosen above.
5. Interpret the confidence interval in the context of the question.
6. Does the confidence interval suggest that the mean SAR exceeds 0.9? Compare your answer with what you got for the hypothesis test. Do the confidence interval and hypothesis test support each other? Explain your answer.

This is completely made-up.

2. Question 2: A hospital is trying to cut down on emergency room wait times. In the past, they have found that the average wait time is 1.4 hours for patients to be called back to be examined. They have implemented a new triage protocol and are interested in seeing if it has changed the amount of time patients must wait before being called back to be examined. An investigation committee randomly surveyed 70 patients. The sample mean wait time was 1.5 hours with a sample standard deviation of 0.5 hours.

   1. What is the variable being studied?

Categorize it. Based on this, what descriptive statistic (mean or proportion) is best for this situation?

2. Use an appropriate eight-step hypothesis to determine if the average wait time for patients to be called back to be examined has changed from 1.4 hours. Use a level of significance of 10%.
3. Is there a level of significance that causes you to change your decision?
4. Suppose the true population mean wait time is 1.4 hours, have you made an error in b)? If so, what type?
5. Construct a 90% confidence interval for the population mean emergency room wait times.
6. Interpret the confidence interval in the context of the question .
7. If the investigation committee wants to increase its level of confidence and keep the margin of error the same by taking another survey, what changes should it make?
8. If the investigation committee did another survey, kept the margin of error the same, and surveyed 200 people instead of 70, how would the level of confidence have to change? Why?
9. Suppose investigation committee wanted their estimate of the population mean emergency room wait times to be within

0.05 hours of the true mean. How many patients would they need to interview?

3. Question 3: Twenty-five Americans were surveyed to determine the number of hours they spend watching television each month. The results were as follows:

207	188	168	122	107
122	173	190	140	129
205	169	163	118	142
150	130	123	129	97
156	118	150	129	216

Assume that the underlying population distribution is normal and the population standard deviation is known to be 32 hours.

1. What is the variable being studied? Categorize it. Based on this, what descriptive statistic (mean or proportion) is best for this situation?
2. The U.S. government has recently released a recommendation that Americans watch less than 150 hours of television per month. Based on this sample, is there enough evidence to suggest that, on average, Americans are meeting this recommendation? Base your answer on an appropriate eight-step hypothesis test. Use $\alpha = 5\%$.
3. Construct a 99% confidence interval for

the population mean hours spent watching television per month.

4. Interpret the confidence interval in the context of the question.
5. Explain what the confidence level means in the context of the question.

4. Question 4: The standard deviation of the weights of newborn elephants is known to be approximately 15 pounds. We wish to construct a 95% confidence interval for the mean weight of newborn elephant calves. Fifty newborn elephants are weighed. The sample mean is 244 pounds. The sample standard deviation is 11 pounds.

   1. What model will you use to construct a confidence interval for the population mean? Explain your reasoning by referring to the criteria for that model.
   2. Construct a 95% confidence interval for the population mean weight of newborn elephants.
   3. What will happen to the confidence interval obtained, if 500 newborn elephants are weighed instead of 50? Why?
   4. Based on the confidence interval, is it fair to say that the average weight of a newborn elephants exceeds 235 pounds? Explain your answer.

    5. Does an appropriate hypothesis test support your decision in d)? Explain your answer by doing the eight-step hypothesis test.

5. Question 5: A news magazine is investigating the changing dynamics in marriages. Historically, men made many of the financial decisions including the decision on whether to make major household purchases (such as buying a new vehicle or doing a renovation), while women were left out of them. To investigate whether this has changed, the magazine is considering doing a study to find out the percentage of couples who are equally involved in making decision about household purchases.

    1. What is the variable being studied? Categorize it. Based on this, what descriptive statistic (mean or proportion) is best for this situation?
    2. When designing a study to determine this population proportion, what is the minimum number you would need to survey to be 90% confident that the population proportion is estimated to within 0.05?
    3. If it were later determined that it was important to be more than 90% confident, how would it affect the minimum number

you need to survey? Why? Do not do any calculations. Suppose the marketing company did do the survey. They randomly surveyed 200 households and found that in 114 of them, the couple makes major household purchasing decisions together. A similar study from the 1980s found that 46.5% of couple made major household purchasing decisions together

4. Conduct an eight-step hypothesis test to determine whether there has been a significant increase in the number of couples who make major household purchasing decisions together since the 1980s. The editor of the magazine will only publish the article if there is ample evidence to support the claim.
5. Construct a 95% confidence interval for the population proportion of couples who make major household purchasing decisions together.
6. Interpret the confidence interval in the context of the question.
7. If the rate has increased, use the confidence interval to determine by how much the rate has increased since the 1980s.
8. List two difficulties the company might have in obtaining random results, if this survey were done by email.

6. Question 6: Suppose that an accounting firm has developed a new software to help their clients do their taxes more quickly. Based off of a national survey, most people spend 24.4 hours completing their personal income taxes a year. The accounting firm has a random sample of 100 of their clients complete their 2016 income tax return using the new software. The sample mean time to complete the tax returns is 23.6 hours with a standard deviation of 7.0 hours. The firm doesn't want to release the software unless they are sure it will reduce the time it takes clients to do their taxes. The population distribution is assumed to be normal.

    1. What is the variable being studied? Categorize it. Based on this, what descriptive statistic (mean or proportion) is best for this situation?
    2. Conduct an appropriate eight-step hypothesis test to determine if, on average, the software has reduced the time it takes clients to do their taxes.
    3. Suppose the truth is that the software does help clients do their taxes faster. Has an error been committed? If so, what type of error is it? Explain your answers.
    4. Construct a 90% confidence interval for the population mean time to complete the tax forms.

   5. Interpret the confidence interval in the context of the question.
   6. Does the confidence interval support the results of the hypothesis test? Explain your answer.
   7. If the firm wished to increase its level of confidence and keep the margin of error the same by taking another survey, what changes should it make? Why?
   8. If the firm did another survey, kept the margin of error the same, and only surveyed 49 people, how would the level of confidence have to change? Why?
   9. Suppose that the firm decided that it needed to be at least 96% confident of the population mean length of time to within one hour. How would the number of people the firm surveys change? Why?

7. Question 7: In 2013, it was determined that 21% of North Americans download music illegally. Public Policy Polling is wondering whether that number has changed. They asked a random sample of adults across North America about their downloading habits. When asked, 512 of the 2247 participants admitted that they have illegally downloaded music.

   1. Has the proportion of North Americans who illegally download music increased since 2013? Conduct an appropriate eight-

   step hypothesis test to support your answer.
2. Create and interpret a 99% confidence interval for the true proportion of North American adults who have illegally downloaded music.
3. This survey was conducted through automated telephone interviews on May 6 and 7 of this year. The margin of error of the survey compensates for sampling error, or natural variability among samples. List some factors that could affect the surveyÕs outcome that are not covered by the margin of error.
4. Without performing any calculations, describe how the confidence interval would change if the confidence level changed from 99% to 90%.
5. Suppose Public Policy Polling want to conduct the study again now. They want to keep the same level of confidence as their last survey, but they want their results to within 2% of the true proportion of Canadian adults who have illegally downloaded music. What is the minimum sample size they need to obtain this?

8. Question 8: A survey of the mean number of cents off that coupons give was conducted by randomly surveying one coupon per page from the coupon sections of a recent San Jose

Mercury News. The following data were collected (in cents): 20; 75 ; 50 ; 65 ; 30 ; 55 ; 40 ; 40 ; 30 ; 55 ; 150; 40 ; 65 ; 40 . Assume the underlying distribution is approximately normal.

1. What is the variable being studied? Categorize it. Based on this, what descriptive statistic (mean or proportion) is best for this situation?
2. Conduct an appropriate eight-step hypothesis test to determine if the mean number of cents off a coupon is different from 50 . Use a level of significance of 3%.
3. What is the probability of committing a type I error in the above hypothesis test?
4. Construct a 97% confidence interval for the population mean worth of coupons.
5. Interpret the confidence interval in the context of the question.
6. If many random samples were taken of size 14, what percent of the confidence intervals constructed should contain the population mean worth of coupons? Explain why.

Solutions to Practice questions

1. 1. The variable is the specific absorption rate. It is quantitative continuous data. The best descriptive statistic for this type of data is

the mean.

2. Since the sample size is less than 30, we can only assume the sampling distribution is normal if the population distribution is close to being normal. Based on the normal curve plot and the empirical rule, it appears that the sample is not normally distributed. The normal curve plot is not a straight line and only 55.6% of the data fall within the first standard deviation of this. This conclusion is supported by a bimodal histogram. This suggests that the population distribution is not normal, which means we cannot be certain the sampling distribution is normal. Regardless of your answer in b), assume that the sampling distribution is normal for the remaining questions.

3. State hypotheses both in sentence and numerical form. Define the symbols. H0: on average, cell phones have SAR rates that are 0.9 watts per kg, $\mu=0.9$; HA: on average, cell phones have SAR rates that exceed 0.9 watts per kg, $\mu>0.9$ State the evidence. $n=27,\bar{X}=0.989,s=0.410$ State the model. Explain why you have chosen it. Therefore, since we need to estimate the population standard deviation using the sample standard deviation and the sample size is small, we will use the t-

based mean model.

- Sampling distribution of sample means is normal? Yes, as stated in the question.
- Population standard deviation is known? No
- Sample size is greater than 40? No

State the level of significance and why you have chosen it. State the decision rule. Since the administrator is risk averse, they want to ensure that they have rejected H0 with a lot of evidence. Therefore, the level of significance that requires the most evidence to reject H0 is 1%. If $p < 1\%$, reject H0. If $p \geq 1\%$, do not reject H0. Evaluate the evidence (i.e. find the p-value using a computer program). $p = 0.1357$ State what the p-value means in the context of the question. The probability that a sample mean SAR of at least 0.989 is observed, under the assumption that the SAR rate is 0.9, is 13.57%.v Make a decision. Since p(13.57%) is greater than α(1%), we do not reject H0. Explain the result in a sentence that refers back to the context. There is not sufficient evidence to suggest that, on average, cell phones have SAR rates that exceed 0.9 watts per kg, which means the advocacy group should not go forward with their petition.

4. Since $\alpha = 1\%$, I will use a confidence level of 98% (for a one-tailed HT, use 1-2*alpha to determine complementary CL): 0.793 to 1.18
5. We are 98% confident that the true population mean for SARs is somewhere between 0.793 watts/kg and 1.18 watts/kg.
6. Though there are possible values for the population mean that do exceed 0.9 watts/kg in the CI, there are also values that do not exceed 0.9 watt/kg. Therefore, the CI would lead to an inconclusive result, meaning it is not clear from the CI whether the pop. mean exceeds 0.9 or not. This aligns with our hypothesis test that there is not enough evidence to suggest that the population mean exceed 0.9 watts/kg.

2.
    1. The variable is the emergency room wait times. It is quantitative continuous data. The best descriptive statistic for this type of data is the mean.

    2. State hypotheses both in sentence and numerical form. Define the symbols. H0: the average wait time for patients to be called back to be examined is 1.4 hours,

$\mu=1.4$; HA: the average wait time for patients to be called back to be examined has changed from 1.4 hours, $\mu\neq1.4$ State the evidence. $n=70,\bar{X}=1.5,s=0.5$ State the model. Explain why you have chosen it. Therefore, since we need to estimate the population standard deviation using the sample standard deviation, but the sample size is large enough that the difference between the t-based and z-based model is minimal, we will use the z-based mean model.

- Sampling distribution of sample means is normal? Yes as the sample size (70) is greater than 30, the central limit theorem applies and the sampling distribution of sample means is normally distributed.
- Population standard deviation is known? No
- Sample size is greater than 40? Yes

State the level of significance and why you have chosen it. State the decision rule. As stated in the question, use 10% If $p<10\%$, reject H0. If $p\geq10\%$, do not reject H0. Evaluate the evidence (i.e. find the p-value using a computer program). $p=0.0943$ State what the p-value means in the context of the question. The probability (times 2) that a sample mean wait time of at least 1.5 hours is observed, under the assumption that the mean wait time is 1.4, is 9.43%. Make a decision. Since p(9.43%) is less than α(10%), we reject H0. Explain the result in a sentence that refers back to the context. There is sufficient evidence to suggest that the average wait time for patients to be called back to be examined has changed from 1.4 hours.

3. Yes, if $\alpha=0.0943$, we would change our decision to do not reject H0.
4. Yes. We have concluded that the mean has changed from 1.4, but the truth is that the mean has stayed the same. Therefore, we have made an error. As we have incorrectly rejected H0 it is a type I error.
5. 1.402 to 1.598
6. We are 90% confident that the population average wait time in the emergency room is somewhere between 1.4 hours and 1.6 hours.
7. If the level of confidence is increased then the critical value in the margin of error would increase. To keep the margin of error the same, either the standard deviation would need to decrease, or the sample size would need to decrease. As the standard deviation is inherent to the data, the sample size needs to decrease.

8. If the sample size increases, then the margin of error decreases. This means that to keep the margin of error constant, the level of confidence would need to increase. This would cause the critical value to be bigger which would compensate for the larger sample size.
9. They would need to interview at least 271 patients.

3.
    1. The variable is the number of hours Americans spend watching TV. It is quantitative discrete data. The best descriptive statistic for this type of data is the mean.

    2. State hypotheses both in sentence and numerical form. Define the symbols. H0: on average, Americans are not meeting this recommendation, $\mu=150$; HA: on average, Americans are meeting this recommendation, $\mu<150$ State the evidence. $n=25,\bar{X}=149.64,\sigma=32$ State the model. Explain why you have chosen it. As the population standard deviation is known, we will use the z-based mean model.

        - Sampling distribution of sample means is normal? Yes.The preamble states the the population is normally

distributed. As the population distribution is assumed to be normal, we know the sampling distribution of sample means is also normal, even though the sample is less than 30.
- Population standard deviation is known? Yes

State the level of significance and why you have chosen it. State the decision rule. The level of significance is provided in the question. If $p < 5\%$, reject H0. If $p \geq 5\%$, do not reject H0. Evaluate the evidence (i.e. find the p-value using a computer program). $p = 0.4776$ State what the p-value means in the context of the question. The probability that a sample mean number of hours of TV watched of at most 149.64 hours is observed, under the assumption that the mean number of hours watching TV is 150, is 47.76%. Make a decision. Since p(47.76%) is greater than α(5%), we do not reject H0. Explain the result in a sentence that refers back to the context. There is not sufficient evidence to suggest that, on average, Americans are meeting the recommendation of watching less than 150 hours of television per month.

3. 133.2 to 166.1

4. We are 99% confident that the population mean time that Americans spend watching TV is somewhere between 133.2 hours and 166.1 hours.
5. The confidence level means that if we took many random samples of size 25 from the population of Americans and constructed many confidence intervals for each of these random samples, then 99% of these confidence intervals will contain the population mean time Americans spend watching TV, while 1% will not.

4.
   1. We know the sampling distribution for sample means is normal because the sample size is greater than 30 as stated in the Central Limit Theorem. Therefore, we use either the Student-t or the standard normal distributions. As the population standard deviation is known, we can use the standard normal distribution (i.e z-based normal distribution).
   2. 239.84 to 248.16
   3. The confidence interval will get narrower because the margin of error will be smaller. The margin of error is smaller because the amount of error between the sample means and the population mean is smaller as stated in the law of large numbers.
   4. Yes, the estimated population mean

weight of newborn elephants is 239.84 pounds to 248.16 pounds. Based on this, it is fair to say that the average weight exceeds 235 pounds, as both bounds are larger than 235.

5. State hypotheses both in sentence and numerical form. Define the symbols. H0: on average, newborn elephants weigh 235 pounds, $\mu = 235$; HA: on average, newborn elephants weigh exceeds 235 pounds, $\mu > 235$ State the evidence. $n = 50, \bar{X} = 244, \sigma = 15$ State the model. Explain why you have chosen it. As the population standard deviation is known, we will use the z-based mean model.

    - Sampling distribution of sample means is normal? Yes as the sample size (50) is greater than 30, the central limit theorem applies and the sampling distribution of sample means is normally distributed.
    - Population standard deviation is known? Yes

    State the level of significance and why you have chosen it. State the decision rule. As the confidence level in the previous question was 95% and we are attempting to verify the CI with a HT, we should use an α of 2.5% (solve for alpha in 0.95 = 1-2*alpha, for a one-tailed HT). If $p < 2.5\%$, reject H0. If $p \geq 2.5\%$, do not reject H0. Evaluate the evidence (i.e. find the p-value using a computer program). p=1.10E-5=1.10×10-5=0.000011 State what the p-value means in the context of the question. The probability that a sample mean weight of newborn elephants is at least 244 pounds is observed, under the assumption that the mean weight of newborn elephants is 235, is 0.0011%. Make a decision. Since p(0.0011%) is less than α(5%), we reject H0. Explain the result in a sentence that refers back to the context. There is sufficient evidence to suggest that, on average, newborn elephants weigh exceeds 235 pounds.

5.
    1. The variable is what whether a couple makes major household purchasing decisions together or not. It is categorical nominal data. The best descriptive statistic for this type of data is a proportion.
    2. They would need to interview a minimum of 271 households (Note: As no estimate of the population proportion is provided, use 50%)
    3. If it were later determined that it was important to be more than 90% confident and a new survey were commissioned,

how would it affect the minimum number you need to survey? Why?

4. State hypotheses both in sentence and numerical form. Define the symbols. H0: the proportion of couples who make major household purchasing decisions together is unchanged at 46.5%, $\pi = 0.465$; HA: the proportion of couples who make major household purchasing decisions together is greater than 46.5%, $\pi > 0.465$ State the evidence. $n = 200, X = 114$ State the model. Explain why you have chosen it.

    - Binomial distribution? Yes, because ...

        - Data is being counted: Yes. Counting the number of couples.
        - Data is collected randomly: Says so in question
        - There are only two outcomes: Either couple makes household decisions together or they don't.
        - There are a fixed number of trials: 200
        - The trials are independent: As the sample is random, it is safe to say this is the case as how one couple makes decisions should not effect how another randomly selected couple makes decisions.

State the level of significance and why you have chosen it. State the decision rule. As the editor needs strong evidence, need to choose α to be small, i.e. 1%. If $p < 1\%$, reject H0. If $p \geq 1\%$, do not reject H0. Evaluate the evidence (i.e. find the p-value using a computer program). $p = 0.00185$ State what the p-value means in the context of the question. The probability that at least 114 out of 200 couples make major purchasing together, assuming the rate has not changed since the 1980s, is 0.19%. Make a decision. Since p(0.19%) is less than α(1%), we reject H0. Explain the result in a sentence that refers back to the context. There is sufficient evidence to suggest that the proportion of couples who make major household purchasing decisions together is greater than 46.5%.

5. 0.5014 to 0.6386
6. We are 95% confident that the true proportion of couples who make major household purchasing decisions together is somewhere between 50.14% and 63.86%.
7. Based off of the CI, the rate has increased by at least 3.6% and by at most 17.4%.
8. One issue is how will the marketing company develop the list of email addresses. Most likely they will not have a complete list of all emails for all

households. Second of all, the email will be sent to a member of the household and not to the household as a whole. Thus one household may get multiple surveys. Further, not everyone uses email so the sample will miss those households.

6. 1. The variable is the amount of time people take completing their tax forms. It is quantitative continuous data. The best descriptive statistic for this type of data is the mean.
   2. Conduct an appropriate eight-step hypothesis test to determine if, on average, the software has reduced the time it takes clients to do their taxes.

      State hypotheses both in sentence and numerical form. Define the symbols. H0: on average, the software has not reduced the time it takes clients to do their taxes, $\mu = 24.4$; HA: on average, the software has reduced the time it takes clients to do their taxes, $\mu < 24.4$ State the evidence. $n = 100, \bar{X} = 23.6, s = 7.0$ State the model. Explain why you have chosen it. Therefore, since we need to estimate the population standard deviation using the sample standard deviation but the sample size is large enough that there the difference between the z-based and t-based models is minimal, we will use the z-based mean model.

- Sampling distribution of sample means is normal? Yes as the population distribution is assumed to be normal, we know the sampling distribution of sample means is also normal.
- Population standard deviation is known? No
- Sample size greater than 40? Yes

State the level of significance and why you have chosen it. State the decision rule. Since the firm doesn't want to release the software unless they are very confident that it works, they should choose a small level of significance (i.e. 1%). If $p < 1\%$, reject H0. If $p \geq 1\%$, do not reject H0. Evaluate the evidence (i.e. find the p-value using a computer program). $p = 0.1265$ State what the p-value means in the context of the question. The probability that a sample mean time to complete tax returns of at most 23.6 hours is observed, under the assumption that the mean time is 24.4, is 12.65%. Make a decision. Since p(12.65%) is greater than α(1%), we do not reject H0. Explain the result in a sentence that refers back to the context.

There is not sufficient evidence to suggest that, on average, the software has reduced the time it takes clients to do their taxes.

3. Since we have stated that it is that there is not enough evidence that the software has reduced the time it takes clients to do their taxes, when in fact it has, we have committed a type II error.
4. 22.45 to 24.75
5. We are 90% confident that the true average time it takes for people to complete their tax forms with this new software is somewhere between 22.45 hours and 24.75 hours.
6. The HT has led us to state that there is evidence that the average time has not been reduced from 24.4. The CI supports this as it contains the population mean of 24.4 hours.
7. If the level of confidence is increased then the critical value in the margin of error would increase. To keep the margin of error the same, either the standard deviation would need to decrease, or the sample size would need to increase. As the standard deviation is inherent to the data, the sample size needs to increase.
8. If the sample size decreases, then the margin of error increases. This means that to keep the margin of error constant, the

level of confidence would need to decrease. This would cause the critical value to be smaller which would compensate for the smaller sample size.

9. It would not change the number of people needed to be interviewed. The level of confidence and the sample size are independent of each other.

7.
    1. State hypotheses both in sentence and numerical form. Define the symbols. H0: the proportion of North Americans who illegally download music not increased since 2013, $\pi=0.21$HA: the proportion of North Americans who illegally download music increased since 2013, $\pi>0.21$ State the evidence. $n=2247, X=512$ State the model. Explain why you have chosen it. Normal distribution approximation of the binomial distribution: Therefore, we will use the normal distribution approximation of the binomial distribution.

        - Binomial distribution? Yes, because ...

            - Data is being counted: Yes. Counting the number of North Americans who illegally download music.
            - Data is collected randomly: Says so in question

- ○ There are only two outcomes: Either person illegally downloads music or they don't.
- ○ There are a fixed number of trials: 2247
- ○ The trials are independent: As the sample is random, it is safe to say this is the case as whether one person downloads music illegally or not should not effect whether another randomly selected person does.

State the level of significance and why you have chosen it. State the decision rule. As there is no motivation stated in the study, I will choose a level of significance that is a balance between rejecting and not rejecting H0, i.e. 5%. If $p<5\%$, reject H0. If $p\geq5\%$, do not reject H0. Evaluate the evidence (i.e. find the p-value using a computer program). $p=0.0208$ State what the p-value means in the context of the question. The probability that at least 512 out of 2247 North Americans admit that they have downloaded music illegally, assuming the rate has not changed since 2013, is 2.08%. Make a decision. Since p(2.08%) is less than α(5%), we reject H0. Explain the result in a sentence that refers back to the context. There is sufficient evidence to suggest that the proportion of North Americans who illegally download music increased since 2013.

2. We are 99% confident that the true proportion of Canadians that download music illegally is somewhere between 20.51% and 25.07%.
3. Some people may not want to admit to having downloaded music illegally. It is unclear how PPP got the list of phone numbers. This list could miss cell phone users and thus would not be representative.
4. The confidence interval would get narrower.
5. 2919

8.
   1. The variable is the number of cents off that coupons give. It is quantitative discrete data. The best descriptive statistic for this type of data is the mean.

   2. State hypotheses both in sentence and numerical form. Define the symbols. H0: the mean number of cents off a coupon is the same as 50 , $\mu = 50$; HA: the mean number of cents off a coupon is different from 50 , $\mu \neq 50$ State the evidence. $n = 14, \bar{X} = 53.93, s = 31.63$ State the model. Explain why you have chosen it.

Therefore, since we need to estimate the population standard deviation using the sample standard deviation and the sample size is small, we will use the t-based mean model.

- Sampling distribution of sample means is normal? Yes as the population distribution is assumed to be normal, we know the sampling distribution of sample means is also normal.
- Population standard deviation is known? No
- Sample size greater than 40? No

State the level of significance and why you have chosen it. State the decision rule. Level of significance in the question is stated to be 3%. If $p < 3\%$, reject H0. If $p \geq 3\%$, do not reject H0. Evaluate the evidence (i.e. find the p-value using a computer program). $p = 0.6499$ State what the p-value means in the context of the question. The probability (times 2) that a sample mean number of cents off a coupon of at most 53.929 is observed, under the assumption that the mean number of cents is 50, is 64.99%. Make a decision. Since p(64.99%) is greater than α(3%), we do not reject H0. Explain the result in a sentence that refers back to the context. There is not sufficient evidence to suggest that the mean number of cents off a coupon is different from 50 .

3. It is the level of significance, 3%.
4. 33.335 to 74.522
5. We are 97% confident that the mean number of cents off that coupons give is somewhere between 33.3 and 74.5 .
6. 97% of them would contain the population mean, while 3% would not. This is determined by the confidence level.

Introduction to Regression

class="introduction"

Introduction to regression

We encounter statistics in our daily lives more often than we probably realize and from many different sources, like the news. (credit: David Sim)

You are probably asking yourself the question, "When and where will I use statistics?" If you read any newspaper, watch television, or use the Internet, you will see statistical information. There are statistics about crime, sports, education, politics, and real estate. Typically, when you read a newspaper article or watch a television news program, you are given sample information. With this information, you may make a decision about the correctness of a statement, claim, or "fact." Statistical methods can help you make the "best educated guess."

Since you will undoubtedly be given statistical information at some point in your life, you need to know some techniques for analyzing the information thoughtfully. Think about buying a house or managing a budget. Think about your chosen profession. The fields of economics, business, psychology, education, biology, law, computer science, police science, and early childhood development require at least one course in statistics.

Included in this chapter are the basic ideas and words of probability and statistics. You will soon understand that statistics and probability work together. You will also learn how data are gathered and what "good" data can be distinguished from "bad."

The Correlation Coefficient r

As we begin this section we note that the type of data we will be working with has changed. Perhaps unnoticed, all the data we have been using is for a single variable. It may be from two samples, but it is still a univariate variable. The type of data described in the examples above and for any model of cause and effect is **bivariate** data — "bi" for two variables. In reality, statisticians use **multivariate** data, meaning many variables.

For our work we can classify data into three broad categories, time series data, cross-section data, and panel data. We met the first two very early on. Time series data measures a single unit of observation; say a person, or a company or a country, as time passes. What are measured will be at least two characteristics, say the person’s income, the quantity of a particular good they buy and the price they paid. This would be three pieces of information in one time period, say 1985. If we followed that person across time we would have those same pieces of information for 1985,1986, 1987, etc. This would constitute a times series data set. If we did this for 10 years we would have 30 pieces of information concerning this person’s consumption habits of this good for the past decade and we would know their income and the price they paid.

A second type of data set is for cross-section data.

Here the variation is not across time for a single unit of observation, but across units of observation during one point in time. For a particular period of time we would gather the price paid, amount purchased, and income of many individual people.

A third type of data set is panel data. Here a panel of units of observation is followed across time. If we take our example from above we might follow 500 people, the unit of observation, through time, ten years, and observe their income, price paid and quantity of the good purchased. If we had 500 people and data for ten years for price, income and quantity purchased we would have 15,000 pieces of information. These types of data sets are very expensive to construct and maintain. They do, however, provide a tremendous amount of information that can be used to answer very important questions. As an example, what is the effect on the labor force participation rate of women as their family of origin, mother and father, age? Or are there differential effects on health outcomes depending upon the age at which a person started smoking? Only panel data can give answers to these and related questions because we must follow multiple people across time. The work we do here however will not be fully appropriate for data sets such as these.

Beginning with a set of data with two independent variables we ask the question: are these related?

One way to visually answer this question is to create a scatter plot of the data. We could not do that before when we were doing descriptive statistics because those data were univariate. Now we have bivariate data so we can plot in two dimensions. Three dimensions are possible on a flat piece of paper, but become very hard to fully conceptualize. Of course, more than three dimensions cannot be graphed although the relationships can be measured mathematically.

To provide mathematical precision to the measurement of what we see we use the correlation coefficient. The correlation tells us something about the co-movement of two variables, but **nothing** about why this movement occurred. Formally, correlation analysis assumes that both variables being analyzed are **independent** variables. This means that neither one causes the movement in the other. Further, it means that neither variable is dependent on the other, or for that matter, on any other variable. Even with these limitations, correlation analysis can yield some interesting results.

The correlation coefficient, ρ (pronounced rho), is the mathematical statistic for a population that provides us with a measurement of the strength of a linear relationship between the two variables. For a sample of data, the statistic, r, developed by Karl Pearson in the early 1900s, is an estimate of the population correlation and is defined mathematically as:

$$r = 1n-1\ \Sigma(X1i - X–1)(X2i - X–2)\ sx1sx2$$

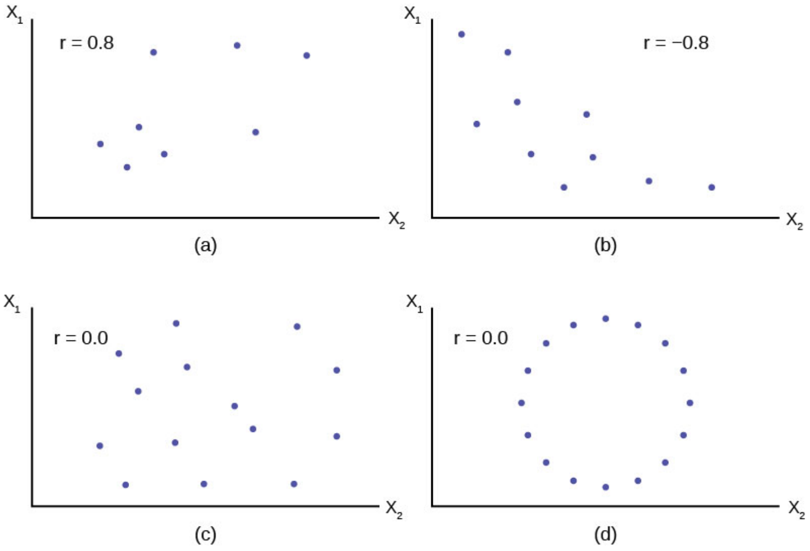
OR

$$r = \Sigma X1iX2i - nX–1 - X–2\ (\Sigma X12i - nX–12)\ (\Sigma X22i - nX–22)$$

where s_{x1} and s_{x2} are the standard deviations of the two independent variables X_1 and X_2, X–1 and X–2 are the sample means of the two variables, and X_{1i} and X_{2i} are the individual observations of X_1 and X_2. The correlation coefficient r ranges in value from -1 to 1. The second equivalent formula is often used because it may be computationally easier. As scary as these formulas look they are really just the ratio of the covariance between the two variables and the product of their two standard deviations. That is to say, it is a measure of relative variances.

In practice all correlation and regression analysis will be provided through computer software designed for these purposes. Anything more than perhaps one-half a dozen observations creates immense computational problems. It was because of this fact that correlation, and even more so, regression, were not widely used research tools until after the advent of “computing machines”. Now the computing power required to analyze data using regression packages is deemed almost trivial by comparison to just a decade ago.

To visualize any **linear** relationship that may exist review the plot of a scatter diagrams of the standardized data. [link] presents several scatter diagrams and the calculated value of r. In panels (a) and (b) notice that the data generally trend together, (a) upward and (b) downward. Panel (a) is an example of a positive correlation and panel (b) is an example of a negative correlation, or relationship. The sign of the correlation coefficient tells us if the relationship is a positive or negative (inverse) one. If all the values of X1 and X2 are on a straight line the correlation coefficient will be either 1 or -1 depending on whether the line has a positive or negative slope and the closer to one or negative one the stronger the relationship between the two variables. BUT ALWAYS REMEMBER THAT THE CORRELATION COEFFICIENT DOES NOT TELL US THE SLOPE.



Remember, all the correlation coefficient tells us is whether or not the data are linearly related. In panel (d) the variables obviously have some type of very specific relationship to each other, but the correlation coefficient is zero, indicating no **linear** relationship exists.

If you suspect a linear relationship between X_1 and X_2 then *r* can measure how strong the linear relationship is.

What the VALUE of *r* tells us:

- The value of *r* is always between –1 and $+1$: –1 $\leq \mathrm{r} \leq 1$.
- The size of the correlation *r* indicates the strength of the **linear** relationship between X_1 and X_2. Values of *r* close to –1 or to $+1$ indicate a stronger linear relationship between X_1 and X_2.
- If $r = 0$ there is absolutely no linear relationship between X_1 and X_2 **(no linear correlation)**.
- If $r = 1$, there is perfect positive correlation. If $r = –1$, there is perfect negative correlation. In both these cases, all of the original data points lie on a straight line: ANY straight line no matter what the slope. Of course, in the real world, this will not generally happen.

What the SIGN of *r* tells us

- A positive value of *r* means that when X1 increases, X2 tends to increase and when X1 decreases, X2 tends to decrease **(positive correlation)**.
- A negative value of *r* means that when X1 increases, X2 tends to decrease and when X1 decreases, X2 tends to increase **(negative correlation)**.

Note

Strong correlation does not suggest that X1 causes X2 or X2 causes X1. We say "**correlation does not imply causation.**"

In order to have a correlation coefficient between traits A and B, it is necessary to have:

1. one group of subjects, some of whom possess characteristics of trait A, the remainder possessing those of trait B
2. measures of trait A on one group of subjects and of trait B on another group
3. two groups of subjects, one which could be classified as A or not A, the other as B or not B

4. two groups of subjects, one which could be classified as A or not A, the other as B or not B

---

d

Define the Correlation Coefficient and give a unique example of its use.

A measure of the degree to which variation of one variable is related to variation in one or more other variables. The most commonly used correlation coefficient indicates the degree to which variation in one variable is described by a straight line relation with another variable.

Suppose that sample information is available on family income and Years of schooling of the head of the household. A correlation coefficient = 0 would indicate no linear association at all between these two variables. A correlation of 1 would indicate perfect linear association (where all variation in family income could be associated with schooling and vice versa).

If the correlation between age of an auto and money spent for repairs is +.90

1. 81% of the variation in the money spent for repairs is explained by the age of the auto
2. 81% of money spent for repairs is unexplained by the age of the auto
3. 90% of the money spent for repairs is explained by the age of the auto
4. none of the above

---

a. 81% of the variation in the money spent for repairs is explained by the age of the auto

Suppose that college grade-point average and verbal portion of an IQ test had a correlation of .40. What percentage of the variance do these two have in common?

1. 20
2. 16
3. 40
4. 80

---

b. 16

True or false? If false, explain why: The coefficient of determination can have values between -1 and +1.

The coefficient of determination is r··2 with 0 ≤ r··2 ≤ 1, since -1 ≤ r ≤ 1.

True or False: Whenever r is calculated on the basis of a sample, the value which we obtain for r is only an estimate of the true correlation coefficient which we would obtain if we calculated it for the entire population.

True

Under a "scatter diagram" there is a notation that the coefficient of correlation is .10. What does this mean?

1. plus and minus 10% from the means includes about 68% of the cases
2. one-tenth of the variance of one variable is shared with the other variable
3. one-tenth of one variable is caused by the other variable
4. on a scale from -1 to +1, the degree of linear relationship between the two variables is +.10

---

d. on a scale from -1 to +1, the degree of linear relationship between the two variables is +.10

The correlation coefficient for X and Y is known to be zero. We then can conclude that:

1. X and Y have standard distributions
2. the variances of X and Y are equal
3. there exists no relationship between X and Y
4. there exists no linear relationship between X and Y
5. none of these

---

d. there exists no linear relationship between X and Y

What would you guess the value of the correlation coefficient to be for the pair of variables: "number of man-hours worked" and "number of units of work completed"?

1. Approximately 0.9
2. Approximately 0.4
3. Approximately 0.0
4. Approximately -0.4
5. Approximately -0.9

Approximately 0.9

In a given group, the correlation between height measured in feet and weight measured in pounds is $+.68$. Which of the following would alter the value of r?

1. height is expressed centimeters.
2. weight is expressed in Kilograms.
3. both of the above will affect r.
4. neither of the above changes will affect r.

---

d. neither of the above changes will affect r.

Glossary

Bivariate

two variables are present in the model where one is the "cause" or independent variable and the other is the "effect" of dependent variable.

Multivariate

a system or model where more than one independent variable is being used to predict an outcome. There can only ever be one dependent variable, but there is no limit to the number of independent variables.

R – Correlation Coefficient

A number between -1 and 1 that represents the strength and direction of the relationship between "X" and "Y." The value for "*r*" will equal 1 or -1 only if all the plotted points form a perfectly straight line.

Linear

a model that takes data and regresses it into a straight line equation.

Testing the Significance of the Correlation Coefficient

The correlation coefficient, *r*, tells us about the strength and direction of the linear relationship between X_1 and X_2.

The sample data are used to compute *r*, the correlation coefficient for the sample. If we had data for the entire population, we could find the population correlation coefficient. But because we have only sample data, we cannot calculate the population correlation coefficient. The sample correlation coefficient, *r*, is our estimate of the unknown population correlation coefficient.

- $\rho =$ population correlation coefficient (unknown)
- $r =$ sample correlation coefficient (known; calculated from sample data)

The hypothesis test lets us decide whether the value of the population correlation coefficient ρ is "close to zero" or "significantly different from zero". We decide this based on the sample correlation coefficient *r* and the sample size *n*.

If the test concludes that the correlation coefficient is significantly different from zero, we say that the correlation coefficient is "significant."

- Conclusion: There is sufficient evidence to conclude that there is a significant linear relationship between X_1 and X_2 because the correlation coefficient is significantly different from zero.
- What the conclusion means: There is a significant linear relationship X_1 and X_2. If the test concludes that the correlation coefficient is not significantly different from zero (it is close to zero), we say that correlation coefficient is "not significant".

Performing the Hypothesis Test

- **Null Hypothesis: *H_0*: $\rho = 0$**
- **Alternate Hypothesis: *H_a*: $\rho \neq 0$**

What the Hypotheses Mean in Words

- **Null Hypothesis *H_0***: The population correlation coefficient IS NOT significantly different from zero. There IS NOT a significant linear relationship (correlation) between X_1 and X_2 in the population.
- **Alternate Hypothesis *H_a***: The population correlation coefficient is significantly different from zero. There is a significant linear relationship (correlation) between X_1 and X_2 in the population.

Drawing a Conclusion

There are two methods of making the decision concerning the hypothesis. The test statistic to test this hypothesis is:

$$tc = r (1-r2)(n-2)$$

OR

$$tc = rn-2\ 1-r2$$

Where the second formula is an equivalent form of the test statistic, n is the sample size and the degrees of freedom are n-2. This is a t-statistic and operates in the same way as other t tests. Calculate the t-value and compare that with the critical value from the t-table at the appropriate degrees of freedom and the level of confidence you wish to maintain. If the calculated value is in the tail then cannot accept the null hypothesis that there is no linear relationship between these two independent random variables. If the calculated t-value is NOT in the tailed then cannot reject the null hypothesis that there is no linear relationship between the two variables.

A quick shorthand way to test correlations is the relationship between the sample size and the correlation. If:

$$|r| \geq 2n$$

then this implies that the correlation between the two variables demonstrates that a linear relationship exists and is statistically significant at approximately the 0.05 level of significance. As the formula indicates, there is an inverse relationship between the sample size and the required correlation for significance of a linear relationship. With only 10 observations, the required correlation for significance is 0.6325, for 30 observations the required correlation for significance decreases to 0.3651 and at 100 observations the required level is only 0.2000.

Correlations may be helpful in visualizing the data, but are not appropriately used to "explain" a relationship between two variables. Perhaps no single statistic is more misused than the correlation coefficient. Citing correlations between health conditions and everything from place of residence to eye color have the effect of implying a cause and effect relationship. This simply cannot be accomplished with a correlation coefficient. The correlation coefficient is, of course, innocent of this misinterpretation. It is the duty of the analyst to use a statistic that is designed to test for cause and effect relationships and report only those results if they are intending to make such a claim. The problem is that passing this more rigorous test is difficult so lazy and/or unscrupulous "researchers" fall back on correlations when they cannot make their case legitimately.

> Define a t Test of a Regression Coefficient, and give a unique example of its use.

Definition:

A t test is obtained by dividing a regression coefficient by its standard error and then comparing the result to critical values for Students' t with Error *df*. It provides a test of the claim that βi = 0 when all other variables have been included in the relevant regression model.

Example:

Suppose that 4 variables are suspected of influencing some response. Suppose that the results of fitting Yi = β0 + β1X1i + β2X2i + β3X3i + β4X4i + ei include:

Variable	Regression coefficient	Standard error of regular coefficient
.5	1	3
.4	2	+2
.02	3	+1
.6	4	-.5

t calculated for variables 1, 2, and 3 would be 5 or larger in absolute value while that for variable 4 would be less than 1. For most significance levels, the hypothesis β1 = 0 would be rejected. But, notice that this is for the case when X2, X3, and X4 have been included in the regression. For most significance levels, the hypothesis β4 = 0 would be continued (retained) for the case where X1, X2, and X3 are in the regression. Often this pattern of results will result in computing another regression involving only X1, X2, X3, and examination of the t ratios produced for that case.

The correlation between scores on a neuroticism test and scores on an anxiety test is high and positive; therefore

1. anxiety causes neuroticism
2. those who score low on one test tend to score high on the other.
3. those who score low on one test tend to score low on the other.
4. no prediction from one test to the other can be meaningfully made.

---

c. those who score low on one test tend to score low on the other.

Linear Equations

Linear regression for two variables is based on a linear equation with one independent variable. The equation has the form:

$$y = a + bx$$

where *a* and *b* are constant numbers.

The variable ***x* is the independent variable, and *y* is the dependent variable.** Another way to think about this equation is a statement of cause and effect. The X variable is the cause and the Y variable is the hypothesized effect. Typically, you choose a value to substitute for the independent variable and then solve for the dependent variable.

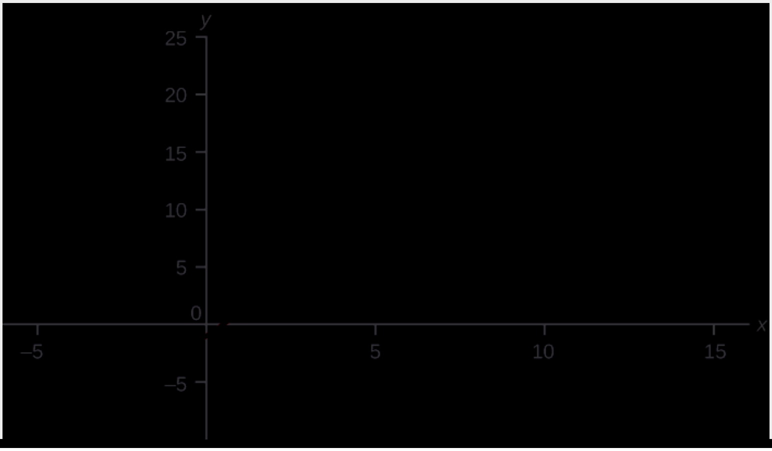
The following examples are linear equations.

$y = 3 + 2x$

$y = -0.01 + 1.2x$

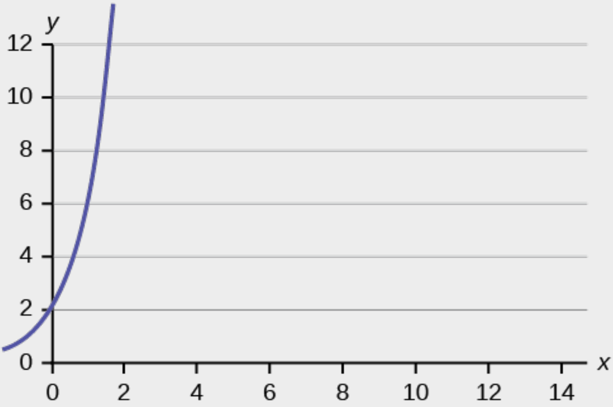
The graph of a linear equation of the form $y = a + bx$ is a **straight line**. Any line that is not vertical can be described by this equation.

Graph the equation $y = -1 + 2x$.



Try It

Is the following an example of a linear equation? Why or why not?



No, the graph is not a straight line; therefore, it is not a linear equation.

Aaron's Word Processing Service (AWPS) does word processing. The rate for services is \$32 per hour plus a \$31.50 one-time charge. The total cost to a customer depends on the number of hours it takes to complete the job.

Find the equation that expresses the **total cost** in terms of the **number of hours** required to complete the job.

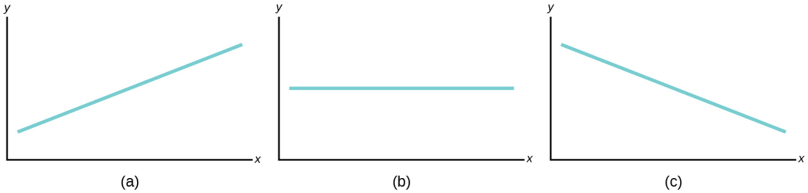
Let $x =$ the number of hours it takes to get the job done.
Let $y =$ the total cost to the customer.

The \$31.50 is a fixed cost. If it takes x hours to complete the job, then $(32)(x)$ is the cost of the word processing only. The total cost is: $y = 31.50 + 32x$

Three possible graphs of $y = a + bx$. (a) If $b > 0$, the line slopes upward to the right. (b) If $b = 0$, the line is horizontal. (c) If $b < 0$, the line slopes downward to the right.

Slope and *Y*-Intercept of a Linear Equation

For the linear equation $y = a + bx$, b = slope and a = y-intercept. From algebra recall that the slope is a number that describes the steepness of a line, and the y-intercept is the y coordinate of the point $(0, a)$ where the line crosses the y-axis. From calculus the slope is the first derivative of the function. For a linear function the slope is $dy / dx = b$ where we can read the mathematical expression as "the change in y (dy) that results from a change in x (dx) $= b * dx$".



Svetlana tutors to make extra money for college. For each tutoring session, she charges a one-time fee of $25 plus $15 per hour of tutoring. A linear equation that expresses the total amount of money Svetlana earns for each session she tutors is $y = 25 + 15x$.

What are the independent and dependent

variables? What is the y-intercept and what is the slope? Interpret them using complete sentences.

The independent variable (x) is the number of hours Svetlana tutors each session. The dependent variable (y) is the amount, in dollars, Svetlana earns for each session.

The y-intercept is 25 ($a = 25$). At the start of the tutoring session, Svetlana charges a one-time fee of $25 (this is when $x = 0$). The slope is 15 ($b = 15$). For each session, Svetlana earns $15 for each hour she tutors.

True or False? If False, correct it: Suppose a 95% confidence interval for the slope β of the straight line regression of Y on X is given by -3.5 $< \beta <$ -0.5. Then a two-sided test of the hypothesis $H0{:}\beta = -1$ would result in rejection of H0 at the 1% level of significance.

False. Since $H0{:}\beta = -1$ would not be rejected at $\alpha = 0.05$, it would not be rejected at $\alpha = 0.01$.

True or False: It is safer to interpret correlation coefficients as measures of association rather than causation because of the possibility of spurious correlation.

True

We are interested in finding the linear relation between the number of widgets purchased at one time and the cost per widget. The following data has been obtained:

X: Number of widgets purchased – 1, 3, 6, 10, 15

Y: Cost per widget(in dollars) – 55, 52, 46, 32, 25

Suppose the regression line is $\hat{y} = -2.5x + 60$. We compute the average price per widget if 30 are purchased and observe which of the following?

1. $\hat{y} = 15 \text{dollars}$; obviously, we are mistaken; the prediction $\hat{y}$ is actually $+15$ dollars.
2. $\hat{y} = 15 \text{dollars}$, which seems reasonable judging by the data.
3. $\hat{y} = -15 \text{dollars}$, which is obvious nonsense. The regression line must be incorrect.

4. $\hat{y} = -15$dollars, which is obvious nonsense. This reminds us that predicting Y outside the range of X values in our data is a very poor practice.

---

d

Discuss briefly the distinction between correlation and causality.

Some variables seem to be related, so that knowing one variable's status allows us to predict the status of the other. This relationship can be measured and is called correlation. However, a high correlation between two variables in no way proves that a cause-and-effect relation exists between them. It is entirely possible that a third factor causes both variables to vary together.

True or False: If r is close to + or -1, we shall say there is a strong correlation, with the tacit understanding that we are referring to a linear relationship and nothing else.

True

Chapter Review

The most basic type of association is a linear association. This type of relationship can be defined algebraically by the equations used, numerically with actual or predicted data values, or graphically from a plotted curve. (Lines are classified as straight curves.) Algebraically, a linear equation typically takes the form $\boldsymbol{y = mx + b}$, where $\boldsymbol{m}$ and $\boldsymbol{b}$ are constants, $\boldsymbol{x}$ is the independent variable, $\boldsymbol{y}$ is the dependent variable. In a statistical context, a linear equation is written in the form $\boldsymbol{y = a + bx}$, where $\boldsymbol{a}$ and $\boldsymbol{b}$ are the constants. This form is used to help readers distinguish the statistical context from the algebraic context. In the equation $y = a + bx$, the constant b that multiplies the $\boldsymbol{x}$ variable (b is called a coefficient) is called as the **slope**. The slope describes the rate of change between the independent and dependent variables; in other words, the rate of change describes the change that occurs in the dependent variable as the independent variable is changed. In the equation $y = a + bx$, the constant a is called as the y-intercept. Graphically, the y-intercept is the y coordinate of the point where the graph of the line crosses the y axis. At this point $x = 0$.

The **slope of a line** is a value that describes the rate of change between the independent and dependent variables. The **slope** tells us how the dependent variable (y) changes for every one unit increase in the independent (x) variable, on average. The ***y*-intercept** is used to describe the dependent variable when the independent variable equals zero. Graphically, the slope is represented by three line types in elementary statistics.

Glossary

Y – the dependent variable

Also, using the letter “y” represents actual values while yˆ represents predicted or estimated values. Predicted values will come from plugging in observed “x” values into a linear model.

X – the independent variable

This will sometimes be referred to as the “predictor” variable, because these values were measured in order to determine what possible outcomes could be predicted.

a is the symbol for the Y-Intercept

Sometimes written as b0, because when writing the theoretical linear model β0 is used to represent a coefficient for a population.

b is the symbol for Slope

The word coefficient will be used regularly for the slope, because it is a number that will always be next to the letter “x.” It will be written as b1 when a sample is used, and β1 will be used with a population or when writing the theoretical linear model.

The Regression Equation

Regression analysis is a statistical technique that can test the hypothesis that a variable is dependent upon one or more other variables. Further, regression analysis can provide an estimate of the magnitude of the impact of a change in one variable on another. This last feature, of course, is all important in predicting future values.

Regression analysis is based upon a functional relationship among variables and further, assumes that the relationship is linear. This linearity assumption is required because, for the most part, the theoretical statistical properties of non-linear estimation are not well worked out yet by the mathematicians and econometricians. This presents us with some difficulties in economic analysis because many of our theoretical models are nonlinear. The marginal cost curve, for example, is decidedly nonlinear as is the total cost function, if we are to believe in the effect of specialization of labor and the Law of Diminishing Marginal Product. There are techniques for overcoming some of these difficulties, exponential and logarithmic transformation of the data for example, but at the outset we must recognize that standard ordinary least squares (OLS) regression analysis will always use a linear function to estimate what might be a nonlinear relationship.

The general linear regression model can be stated by the equation:

$$y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i$$

where β_0 is the intercept, β_i's are the slope between Y and the appropriate X_i, and ε (pronounced epsilon), is the error term that captures errors in measurement of Y and the effect on Y of any variables missing from the equation that would contribute to explaining variations in Y. This equation is the theoretical population equation and therefore uses Greek letters. The equation we will estimate will have the Roman equivalent symbols. This is parallel to how we kept track of the population parameters and sample parameters before. The symbol for the population mean was μ and for the sample mean X– and for the population standard deviation was σ and for the sample standard deviation was s. The equation that will be estimated with a sample of data for two independent variables will thus be:

$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + e_i$$

As with our earlier work with probability distributions, this model works only if certain assumptions hold. These are that the Y is normally distributed, the errors are also normally distributed with a mean of zero and a constant standard deviation, and that the error terms are independent of the size of X and independent of each other.

Assumptions of the Ordinary Least Squares Regression Model

Each of these assumptions needs a bit more explanation. If one of these assumptions fails to be true, then it will have an effect on the quality of the estimates. Some of the failures of these assumptions can be fixed while others result in estimates that quite simply provide no insight into the questions the model is trying to answer or worse, give biased estimates.

1. The independent variables, xi , are all measured without error, and are fixed numbers that are independent of the error term. This assumption is saying in effect that Y is deterministic, the result of a fixed component “X” and a random error component “ϵ.”
2. The error term is a random variable with a mean of zero and a constant variance. The meaning of this is that the variances of the independent variables are independent of the value of the variable. Consider the relationship between personal income and the quantity of a good purchased as an example of a case where the variance is dependent upon the value of the independent variable, income. It is plausible that as income increases the variation around the amount purchased will also increase simply because of the flexibility provided with higher levels of income. The assumption is for

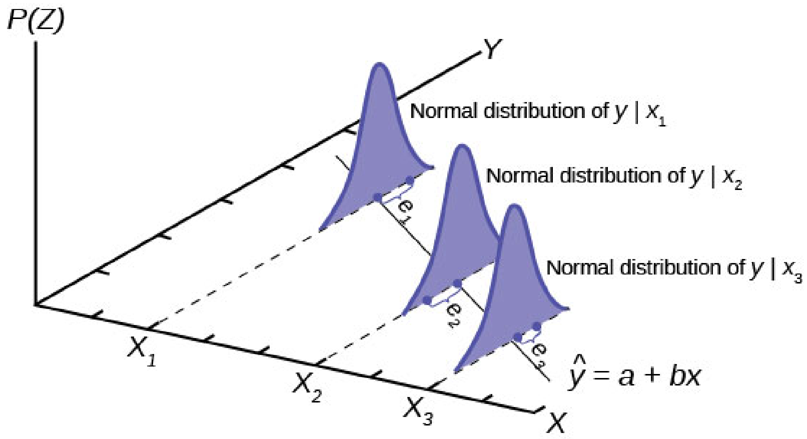
constant variance with respect to the magnitude of the independent variable called homoscedasticity. If the assumption fails, then it is called heteroscedasticity. [link] shows the case of homoscedasticity where all three distributions have the same variance around the predicted value of Y regardless of the magnitude of X.

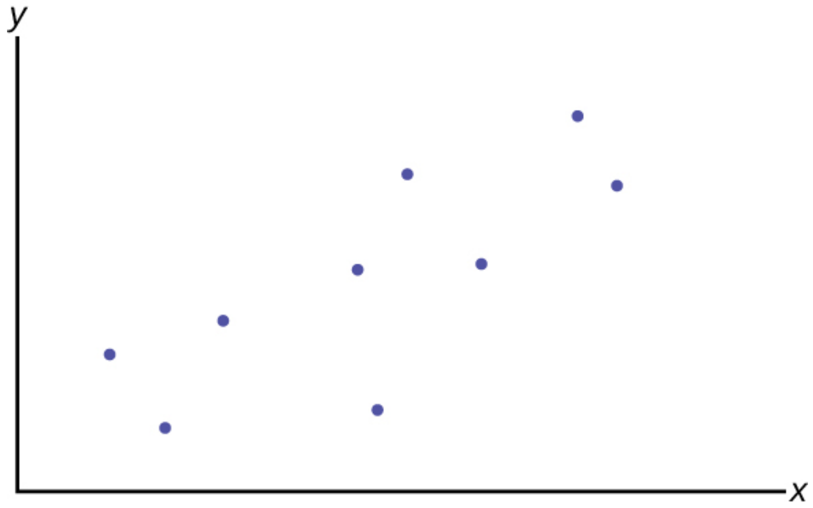
3. While the independent variables are all fixed values they are from a probability distribution that is normally distributed. This can be seen in [link] by the shape of the distributions placed on the predicted line at the expected value of the relevant value of Y.
4. The independent variables are independent of Y, but are also assumed to be independent of the other X variables. The model is designed to estimate the effects of independent variables on some dependent variable in accordance with a proposed theory. The case where some or more of the independent variables are correlated is not unusual. There may be no cause and effect relationship among the independent variables, but nevertheless they move together. Take the case of a simple supply curve where quantity supplied is theoretically related to the price of the product and the prices of inputs. There may be multiple inputs that may over time move together from general inflationary pressure. The input prices will therefore violate this assumption of regression analysis. This

condition is called multicollinearity, which will be taken up in detail later.

5. The error terms are uncorrelated with each other. This situation arises from an effect on one error term from another error term. While not exclusively a time series problem, it is here that we most often see this case. An X variable in time period one has an effect on the Y variable, but this effect then has an effect in the next time period. This effect gives rise to a relationship among the error terms. This case is called autocorrelation, “self-correlated.” The error terms are now not independent of each other, but rather have their own effect on subsequent error terms.

[link] shows the case where the assumptions of the regression model are being satisfied. The estimated line is $y\hat{}=a+bx$. Three values of X are shown. A normal distribution is placed at each point where X equals the estimated line and the associated error at each value of Y. Notice that the three distributions are normally distributed around the point on the line, and further, the variation, variance, around the predicted value is constant indicating homoscedasticity from assumption 2. [link] does not show all the assumptions of the regression model, but it helps visualize these important ones.


P(Z)
Y
Normal distribution of $y \mid x_1$
Normal distribution of $y \mid x_2$
Normal distribution of $y \mid x_3$
e_1
e_2
e_3
$\hat{y} = a + bx$
X_1
X_2
X_3
X



$$y = \beta_0 + \beta_1 X + \varepsilon$$

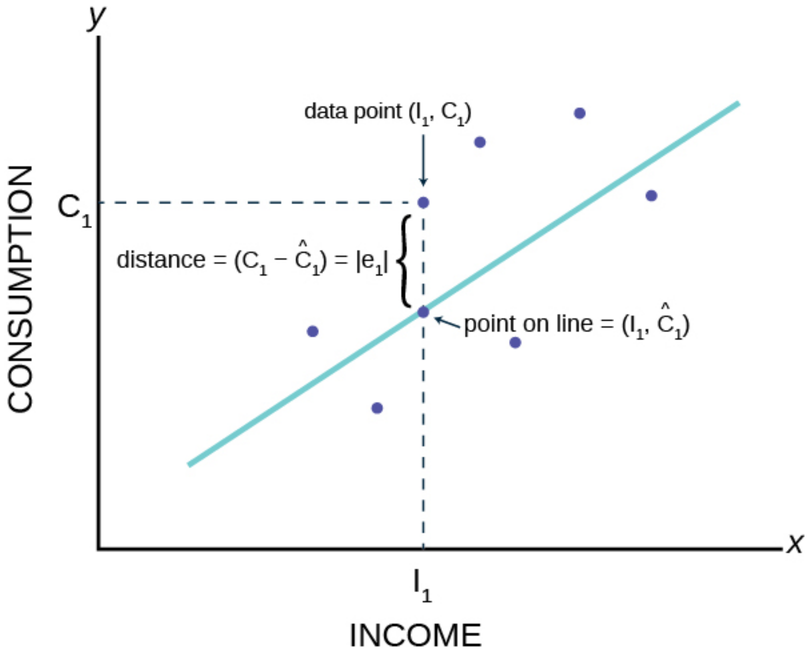
This is the general form that is most often called the multiple regression model. So-called "simple" regression analysis has only one independent (right-hand) variable rather than many independent variables. Simple regression is just a special case of multiple regression. There is some value in beginning with simple regression: it is easy to graph in two dimensions, difficult to graph in three dimensions, and impossible to graph in more than three dimensions. Consequently, our graphs will be for the simple regression case. [link] presents the regression problem in the form of a scatter plot graph of the data set where it is hypothesized that Y is dependent upon the single independent variable X.

A basic relationship from Macroeconomic Principles is the consumption function. This theoretical relationship states that as a person's income rises, their consumption rises, but by a smaller amount than the rise in income. If Y is consumption and X is income in the equation below [link], the regression problem is, first, to establish that this relationship exists, and second, to determine the impact of a change in income on a person's consumption. The parameter β_1 was called the Marginal Propensity to Consume in Macroeconomics Principles.

Each "dot" in [link] represents the consumption and income of different individuals at some point in time. This was called cross-section data earlier; observations on variables at one point in time across different people or other units of measurement. This analysis is often done with time series data, which would be the consumption and income of one individual or country at different points in time. For macroeconomic problems it is common to use times series aggregated data for a whole country. For this particular theoretical concept these data are readily available in the annual report of the President’s Council of Economic Advisors.

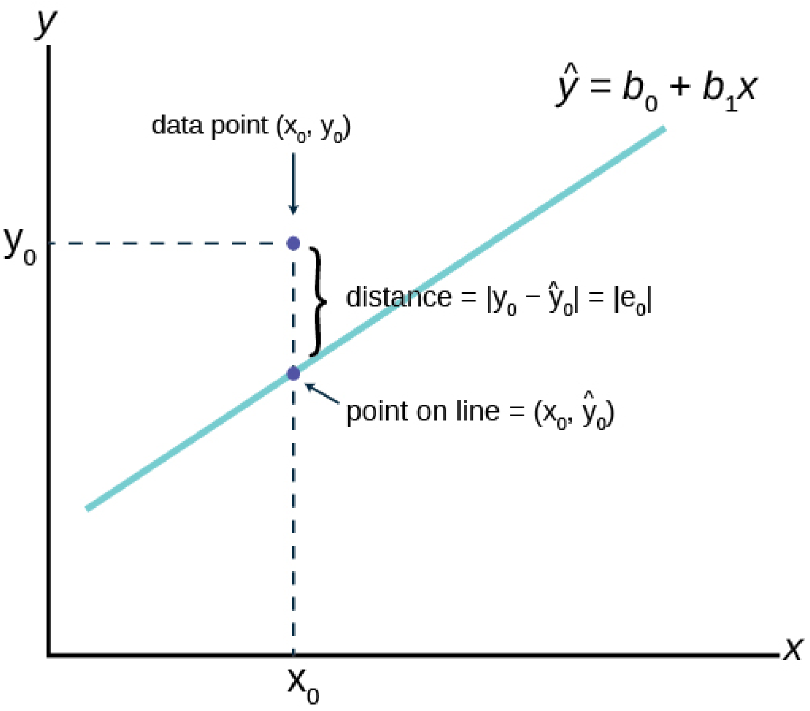
The regression problem comes down to determining which straight line would best represent the data in [link]. Regression analysis is sometimes called "least squares" analysis because the method of determining which line best "fits" the data is to minimize the sum of the squared residuals of a line put through the data.

Population Equation: $C = \beta_0 + \beta_1 \text{ Income} + \varepsilon$
Estimated Equation: $C = b_0 + b_1 \text{ Income} + e$


y
data point (I_1, C_1)
C_1
distance = $(C_1 - \hat{C}_1) = |e_1|$
point on line = $(I_1, \hat{C}_1)$
CONSUMPTION
x
I_1
INCOME

This figure shows the assumed relationship between consumption and income from macroeconomic theory. Here the data are plotted as a scatter plot and an estimated straight line has been drawn. From this graph we can see an error term, e1. Each data point also has an error term. Again, the error term is put into the equation to capture effects on consumption that are not caused by income changes. Such other effects might be a person's savings or wealth, or periods of unemployment. We will see how by minimizing the sum of these errors we can get an estimate for the slope and intercept of this line.

Consider the graph below. The notation has returned to that for the more general model rather than the specific case of the Macroeconomic consumption function in our example.



y
$\hat{y} = b_0 + b_1x$
data point (x_0, y_0)
y_0
distance = $|y_0 - \hat{y}_0| = |e_0|$
point on line = $(x_0, \hat{y}_0)$
x
x_0

The ŷ is read "***y* hat**" and is the **estimated value of *y***. (In [link] Cˆ represents the estimated value of consumption because it is on the estimated line.) It is the value of *y* obtained using the regression line. ŷ is not generally equal to *y* from the data.

The term y0-ŷ0 = e0 is called the **"error" or residual**. It is not an error in the sense of a mistake. The error term was put into the estimating equation to capture missing variables and errors in measurement that may have occurred in the dependent variables. The **absolute value of a residual** measures the vertical distance between the actual value of *y* and the estimated value of *y*. In other words, it measures the vertical distance between the actual data point and the predicted point on the line as can be seen on the graph at point X_0.

If the observed data point lies above the line, the residual is positive, and the line underestimates the actual data value for *y*.

If the observed data point lies below the line, the residual is negative, and the line overestimates that actual data value for *y*.

In the graph, $y0-\hat{y}0 = e0$ is the residual for the point shown. Here the point lies above the line and the residual is positive. For each data point the residuals, or errors, are calculated $y_i – \hat{y}_i = e_i$ for i = 1, 2, 3, ..., n where n is the sample size. Each |e| is a vertical distance.

The sum of the errors squared is the term obviously called **Sum of Squared Errors (SSE)**.

Using calculus, you can determine the straight line that has the parameter values of b_0 and b_1 that minimizes the **SSE**. When you make the **SSE** a minimum, you have determined the points that are on the line of best fit. It turns out that the line of best fit has the equation:
$\hat{y} = b0 + b1\ x$

where b0 = y– − b1 x– and b1 = Σ(x−x–) (y−y–)

$\Sigma(x - x–)2 = cov(x,y)sx2$

The sample means of the *x* values and the *y* values are x– and y–, respectively. The best fit line always passes through the point (x–, y–) called the points of means.

The slope *b* can also be written as:
$b\ 1 = r\ y,x\ (\ s\ y\ s\ x\ )$

where s_y = the standard deviation of the *y* values and s_x = the standard deviation of the *x* values and *r* is the correlation coefficient between *x* and *y*.

These equations are called the Normal Equations and come from another very important mathematical finding called the Gauss-Markov Theorem without which we could not do regression analysis. The Gauss-Markov Theorem tells us that the estimates we get from using the ordinary least squares (OLS) regression method will result in estimates that have some very important properties. In the Gauss-Markov Theorem it was proved that a least squares line is BLUE, which is, **B**est, **L**inear, **U**nbiased, **E**stimator. Best is the statistical property that an estimator is the one with the minimum variance. Linear refers to the property of the type of line being estimated. An unbiased estimator is one whose estimating function has an expected mean equal to the mean of the population. (You will remember that the expected value of μx– was equal to the population mean μ in accordance with the Central Limit Theorem. This is exactly the same concept here).

Both Gauss and Markov were giants in the field of mathematics, and Gauss in physics too, in the 18th century and early 19th century. They barely overlapped chronologically and never in geography, but Markov's work on this theorem was based extensively on the earlier work of Carl Gauss. The extensive applied value of this theorem had to wait until the middle of this last century.

Using the OLS method we can now find the **estimate of the error variance** which is the variance of the squared errors, e^2. This is sometimes called the **standard error of the estimate**. (Grammatically this is probably best said as the estimate of the **error's** variance) The formula for the estimate of the error variance is:

$$s_e^2 = \Sigma (y_i - \hat{y}_i)^2 n - k = \Sigma e_i^2 n - k$$

where ŷ is the predicted value of y and y is the observed value, and thus the term $(y_i - \hat{y}_i)^2$ is the squared errors that are to be minimized to find the estimates of the regression line parameters. This is really just the variance of the error terms and follows our regular variance formula. One important note is that here we are dividing by (n-k), which is the degrees of freedom. The degrees of freedom of a regression equation will be the number of observations, n, reduced by the number of estimated parameters, which includes the intercept as a parameter.

The variance of the errors is fundamental in testing hypotheses for a regression. It tells us just how "tight" the dispersion is about the line. As we will see shortly, the greater the dispersion about the line, meaning the larger the variance of the errors, the less probable that the hypothesized independent variable will be found to have a significant effect on the dependent variable. In short, the theory being tested will more likely fail if the variance of the error term is high. Upon reflection this should not be a surprise. As we tested hypotheses about a mean we observed that large variances reduced the calculated test statistic and thus it failed to reach the tail of the distribution. In those cases, the null hypotheses could not be rejected. If we cannot reject the null hypothesis in a regression problem, we must conclude that the hypothesized independent variable has no effect on the dependent variable.

A way to visualize this concept is to draw two scatter plots of x and y data along a predetermined line. The first will have little variance of the errors, meaning that all the data points will move close to the line. Now do the same except the data points will have a large estimate of the error variance, meaning that the data points are scattered widely along the line. Clearly the confidence about a relationship between x and y is effected by this difference between the estimate of the error variance.

Testing the Parameters of the Line

The whole goal of the regression analysis was to test the hypothesis that the dependent variable, Y, was in fact dependent upon the values of the independent variables as asserted by some foundation theory, such as the consumption function example. Looking at the estimated equation under [link], we see that this amounts to determining the values of b_0 and b_1. Notice that again we are using the convention of Greek letters for the population parameters and Roman letters for their estimates.

The regression analysis output provided by the computer software will produce an estimate of b_0 and b_1, and any other b's for other independent variables that were included in the estimated equation. The issue is how good are these estimates? In order to test a hypothesis concerning any estimate, we have found that we need to know the underlying sampling distribution. It should come as no surprise at his stage in the course that the answer is going to be the normal distribution. This can be seen by remembering the assumption that the error term in the population, ε, is normally distributed. If the error term is normally distributed and the variance of the estimates of the equation parameters, b_0 and b_1, are determined by the variance of the error term, it follows that the variances of the parameter estimates are also normally distributed. And indeed this is just the case.

We can see this by the creation of the test statistic for the test of hypothesis for the slope parameter, β_1 in our consumption function equation. To test whether or not Y does indeed depend upon X, or in our example, that consumption depends upon income, we need only test the hypothesis that β_1 equals zero. This hypothesis would be stated formally as:

$H0:\beta1 = 0$

$Ha:\beta1 \neq 0$

If we cannot reject the null hypothesis, we must conclude that our theory has no validity. If we cannot reject the null hypothesis that $\beta_1 = 0$ then b_1, the coefficient of Income, is zero and zero times anything is zero. Therefore the effect of Income on Consumption is zero. There is no relationship as our theory had suggested.

Notice that we have set up the presumption, the null hypothesis, as "no relationship". This puts the burden of proof on the alternative hypothesis. In other words, if we are to validate our claim of finding a relationship, we must do so with a level of significance greater than 90, 95, or 99 percent. The *status quo* is ignorance, no relationship exists, and to be able to make the claim that we have actually added to our body of knowledge we must do so with significant probability of being correct. John Maynard Keynes got it right and thus was born Keynesian economics starting with this basic concept in 1936.

The test statistic for this test comes directly from our old friend the standardizing formula:

$$t_c = \frac{b_1 - \beta_1}{S_{b_1}}$$

where b_1 is the estimated value of the slope of the regression line, β_1 is the hypothesized value of beta, in this case zero, and S_{b_1} is the standard deviation of the estimate of b_1. In this case we are asking how many standard deviations is the estimated slope away from the hypothesized slope. This is exactly the same question we asked before with respect to a hypothesis about a mean: how many standard deviations is the estimated mean, the sample mean, from the hypothesized mean?

The test statistic is written as a student's t distribution, but if the sample size is larger enough so that the degrees of freedom are greater than 30 we may again use the normal distribution. To see why we can use the student's t or normal distribution we have only to look at S_{b_1} ,the formula for the standard deviation of the estimate of b_1:

$S b 1 = S e 2 (x i - x –) 2$

or

$S b 1 = S e 2 (n - 1) S x 2$

Where S_e is the estimate of the error variance and S_{2x} is the variance of x values of the coefficient of the independent variable being tested.

We see that S_e, the **estimate of the error variance**, is part of the computation. Because the estimate of the error variance is based on the assumption of normality of the error terms, we can conclude that the sampling distribution of the b's, the coefficients of our hypothesized regression line, are also normally distributed.

One last note concerns the degrees of freedom of the test statistic, ν=n-k. Previously we subtracted 1 from the sample size to determine the degrees of freedom in a student's t problem. Here we must subtract one degree of freedom for each parameter estimated in the equation. For the example of the consumption function we lose 2 degrees of freedom, one for b0, the intercept, and one for b_1, the slope of the consumption function. The degrees of freedom would be n - k - 1, where k is the number of independent variables and the extra one is lost because of the intercept. If we were estimating an equation with three independent variables, we would lose 4 degrees of freedom: three for the independent variables, k, and one more for the intercept.

The decision rule for acceptance or rejection of the null hypothesis follows exactly the same form as in all our previous test of hypothesis. Namely, if the calculated value of t (or Z) falls into the tails of the distribution, where the tails are defined by α ,the required significance level in the test, we cannot accept the null hypothesis. If on the other hand, the calculated value of the test statistic is within the critical region, we cannot reject the null hypothesis.

If we conclude that we cannot accept the null hypothesis, we are able to state with (1-α) level of confidence that the slope of the line is given by b_1. This is an extremely important conclusion. Regression analysis not only allows us to test if a cause and effect relationship exists, we can also determine the magnitude of that relationship, if one is found to exist. It is this feature of regression analysis that makes it so valuable. If models can be developed that have statistical validity, we are then able to simulate the effects of changes in variables that may be under our control with some degree of probability , of course. For example, if advertising is demonstrated to effect sales, we can determine the effects of changing the advertising budget and decide if the increased sales are worth the added expense.

Multicollinearity

Our discussion earlier indicated that like all statistical models, the OLS regression model has important assumptions attached. Each assumption, if violated, has an effect on the ability of the model to provide useful and meaningful estimates. The Gauss-Markov Theorem has assured us that the OLS estimates are unbiased and minimum variance, but this is true only under the assumptions of the model. Here we will look at the effects on OLS estimates if the independent variables are correlated. The other assumptions and the methods to mitigate the difficulties they pose if they are found to be violated are examined in Econometrics courses. We take up multicollinearity because it is so often prevalent in Economic models and it often leads to frustrating results.

The OLS model assumes that all the independent variables are independent of each other. This assumption is easy to test for a particular sample of data with simple correlation coefficients. Correlation, like much in statistics, is a matter of degree: a little is not good, and a lot is terrible.

The goal of the regression technique is to tease out the independent impacts of each of a set of independent variables on some hypothesized dependent variable. If two 2 independent variables are interrelated, that is, correlated, then we cannot isolate the effects on Y of one from the other. In an extreme case where x1 is a linear combination of x2, correlation equal to one, both variables move in identical ways with Y. In this case it is impossible to determine the variable that is the true cause of the effect on Y. (If the two variables were actually perfectly correlated, then mathematically no regression results could actually be calculated.)

The normal equations for the coefficients show the effects of multicollinearity on the coefficients.
b1= sy(rx1y - rx1x2 rx2y) sx1 (1- r x1x2 2)
b2= sy(rx2y - rx1x2 rx1y) sx2 (1- r x1x2 2)
b0= y--b1x-1-b2x-2

The correlation between x1 and x2, rx1x22 , appears in the denominator of both the estimating formula for b1 and b2. If the assumption of independence holds, then this term is zero. This indicates that there is no effect of the correlation on the coefficient. On the other hand, as the correlation between the two independent variables increases the denominator decreases, and thus the estimate of the coefficient increases. The correlation has the same effect on both of the coefficients of these two variables. In essence, each variable is “taking” part of the effect on Y that should be attributed to the collinear variable. This results in biased estimates.

Multicollinearity has a further deleterious impact on the OLS estimates. The correlation between the two independent variables also shows up in the formulas for the estimate of the variance for the coefficients.
sb12 = se2 (n-1)sx12(1- rx1x22)
sb22 = se2 (n-1)sx22(1- rx1x22)

Here again we see the correlation between x1 and x2 in the denominator of the estimates of the variance for the coefficients for both variables. If the correlation is zero as assumed in the regression model, then the formula collapses to the familiar ratio of the variance of the errors to the variance of the relevant independent variable. If however the two independent variables are correlated, then the variance of the estimate of the coefficient increases. This results in a smaller t-value for the test of hypothesis of the coefficient. In short, multicollinearity results in failing to reject the null hypothesis that the X variable has no impact on Y when in fact X does have a statistically significant impact on Y. Said another way, the large standard errors of the estimated coefficient created by multicollinearity suggest statistical insignificance even when the hypothesized relationship is strong.

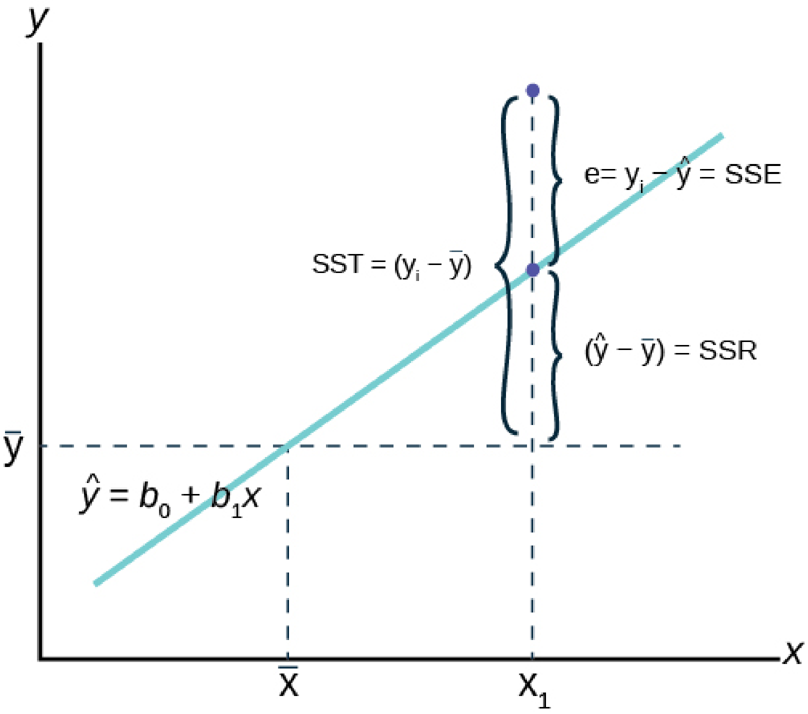
How Good is the Equation?

In the last section we concerned ourselves with testing the hypothesis that the dependent variable did indeed depend upon the hypothesized independent variable or variables. It may be that we find an independent variable that has some effect on the dependent variable, but it may not be the only one, and it may not even be the most important one. Remember that the error term was placed in the model to capture the effects of any missing independent variables. It follows that the error term may be used to give a measure of the "goodness of fit" of the equation taken as a whole in explaining the variation of the dependent variable, Y.

The **multiple correlation coefficient**, also called the **coefficient of multiple determination** or the **coefficient of determination**, is given by the formula:

$$R2 = SSR \ SST$$

where SSR is the regression sum of squares, the squared deviation of the predicted value of y from the mean value of y($\hat{y} - y$–), and SST is the total sum of squares which is the total squared deviation of the dependent variable, y, from its mean value, including the error term, SSE, the sum of squared errors. [link] shows how the total deviation of the dependent variable, y, is partitioned into these two pieces.



[link] shows the estimated regression line and a single observation, x1. Regression analysis tries to explain the variation of the data about the mean value of the dependent variable, y. The question is, why do the observations of y vary from the average level of y? The value of y at observation x1 varies from the mean of y by the difference (yi − y–). The sum of these differences squared is SST, the sum of squares total. The actual value of y at x1 deviates from the estimated value, ŷ, by the difference between the estimated value and the actual value, (yi − ŷ). We recall that this is the error term, e, and the sum of these errors is SSE, sum of squared errors. The deviation of the predicted value of y, ŷ,

from the mean value of y is $(\hat{y} - y\text{–})$ and is the SSR, sum of squares regression. It is called “regression” because it is the deviation explained by the regression. (Sometimes the SSR is called SSM for sum of squares mean because it measures the deviation from the mean value of the dependent variable, y, as shown on the graph.).

Because the SST = SSR + SSE we see that the multiple correlation coefficient is the percent of the variance, or deviation in y from its mean value, that is explained by the equation when taken as a whole. R2 will vary between zero and 1, with zero indicating that none of the variation in y was explained by the equation and a value of 1 indicating that 100% of the variation in y was explained by the equation. For time series studies expect a high R2 and for cross-section data expect low R2.

While a high R2 is desirable, remember that it is the tests of the hypothesis concerning the existence of a relationship between a set of independent variables and a particular dependent variable that was the motivating factor in using the regression model. It is validating a cause and effect relationship developed by some theory that is the true reason that we chose the regression analysis. Increasing the number of independent variables will have the effect of increasing R2. To account for this effect the proper measure of the coefficient of determination is the R–

2, adjusted for degrees of freedom, to keep down mindless addition of independent variables.

There is no statistical test for the R2 and thus little can be said about the model using R2 with our characteristic confidence level. Two models that have the same size of SSE, that is sum of squared errors, may have very different R2 if the competing models have different SST, total sum of squared deviations. The goodness of fit of the two models is the same; they both have the same sum of squares unexplained, errors squared, but because of the larger total sum of squares on one of the models the R2 differs. Again, the real value of regression as a tool is to examine hypotheses developed from a model that predicts certain relationships among the variables. These are tests of hypotheses on the coefficients of the model and not a game of maximizing R2.

Another way to test the general quality of the overall model is to test the coefficients as a group rather than independently. Because this is multiple regression (more than one X), we use the F-test to determine if our coefficients collectively affect Y. The hypothesis is:

Ho:β1 = β2 = … = βi = 0

Ha: "at least one of the βi is not equal to 0"

If the null hypothesis cannot be rejected, then we conclude that none of the independent variables contribute to explaining the variation in Y. Reviewing [link] we see that SSR, the explained sum of squares, is a measure of just how much of the variation in Y is explained by all the variables in the model. SSE, the sum of the errors squared, measures just how much is unexplained. It follows that the ratio of these two can provide us with a statistical test of the model as a whole. Remembering that the F distribution is a ratio of Chi squared distributions and that variances are distributed according to Chi Squared, and the sum of squared errors and the sum of squares are both variances, we have the test statistic for this hypothesis as:

$$Fc= \ (\ SSR \ k \ ) \ (\ SSE \ n-k-1 \ )$$

where *n* is the number of observations and *k* is the number of independent variables. It can be shown that this is equivalent to:

$$Fc \ = \ n-k-1 \ k \cdot R2 \ 1-R2$$

[link] where R2 is the coefficient of determination which is also a measure of the “goodness” of the model.

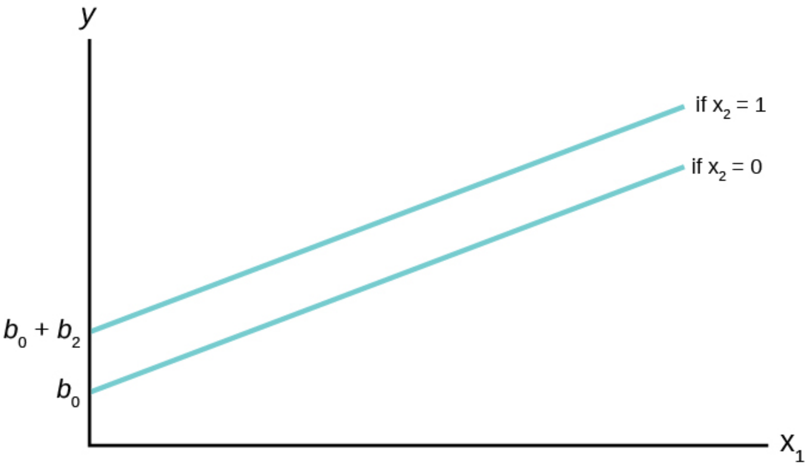
As with all our tests of hypothesis, we reach a conclusion by comparing the calculated F statistic with the critical value given our desired level of confidence. If the calculated test statistic, an F statistic in this case, is in the tail of the distribution, then we cannot accept the null hypothesis. By not being able to accept the null hypotheses we conclude that this specification of this model has validity, because at least one of the estimated coefficients is significantly different from zero.

An alternative way to reach this conclusion is to use the p-value comparison rule. The p-value is the area in the tail, given the calculated F statistic. In essence, the computer is finding the F value in the table for us. The computer regression output for the calculated F statistic is typically found in the ANOVA table section labeled “significance F". How to read the output of an Excel regression is presented below. This is the probability of NOT accepting a false null hypothesis. If this probability is less than our pre-determined alpha error, then the conclusion is that we cannot accept the null hypothesis.

Dummy Variables

Thus far the analysis of the OLS regression technique assumed that the independent variables in the models tested were continuous random variables. There are, however, no restrictions in the regression model against independent variables that are binary. This opens the regression model for testing hypotheses concerning categorical variables such as gender, race, region of the country, before a certain data, after a certain date and innumerable others. These categorical variables take on only two values, 1 and 0, success or failure, from the binomial probability distribution. The form of the equation becomes:

$$\hat{y} = b0 + b2x2 + b1x1$$



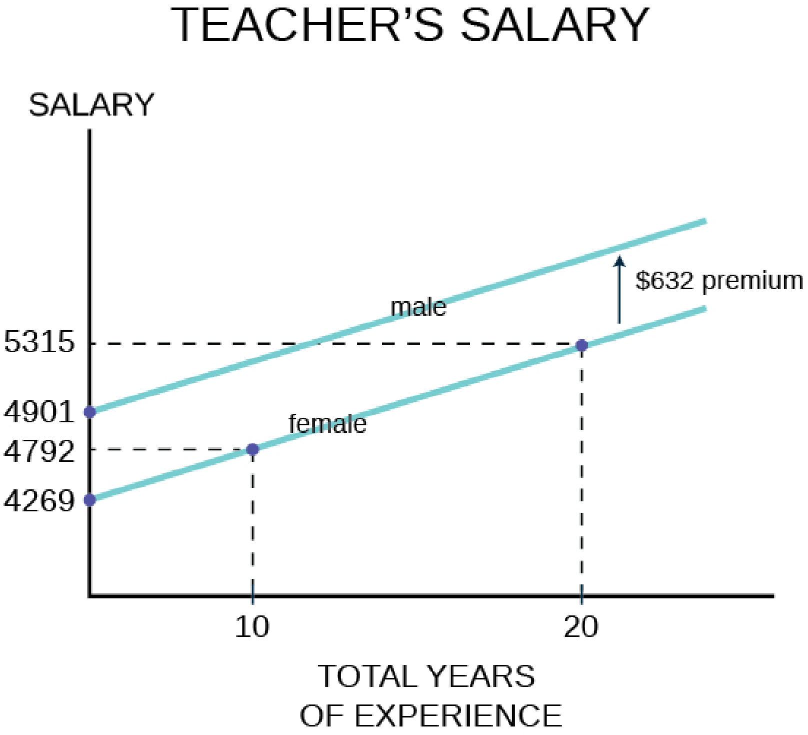
where $x2 = 0{,}1$. X_2 is the dummy variable and X_1 is some continuous random variable. The constant, b_0, is the y-intercept, the value where the line crosses the y-axis. When the value of $X_2 = 0$, the estimated line crosses at b_0. When the value of $X_2 = 1$ then the estimated line crosses at $b_0 + b_2$. In effect the dummy variable causes the estimated line to shift either up or down by the size of the effect of the characteristic captured by the dummy variable. Note that this is a simple parallel shift and does not affect the impact of the other independent variable; X_1.This variable is a continuous random variable and predicts different values of y at different values of X_1 holding constant the condition of the dummy variable.

An example of the use of a dummy variable is the work estimating the impact of gender on salaries. There is a full body of literature on this topic and dummy variables are used extensively. For this example the salaries of elementary and secondary school teachers for a particular state is examined. Using a homogeneous job category, school teachers, and for a single state reduces many of the variations that naturally effect salaries such as differential physical risk, cost of living in a particular state, and other working conditions. The estimating equation in its simplest form specifies salary as a function of various teacher characteristic that economic theory would suggest could affect salary. These would include education level as a measure of potential productivity, age and/or experience to capture on-the-job training, again as a measure of productivity. Because the data are for school teachers employed in a public school districts rather than workers in a for-profit company, the school district's average revenue per average daily student attendance is included as a measure of ability to pay. The results of the regression analysis using data on 24,916 school teachers are presented below.

Variable	Regression Coefficients (b)	Standard Errors of the estimates for teacher's earnings function (s^b)
Intercept	4269.9	
Gender (male = 1)	632.38	13.39
Total Years of Experience	52.32	1.10
Years of Experience in Current District	29.97	1.52
Education	629.33	13.16
Total Revenue per ADA	90.24	3.76
R 2	.725	
n	24,916	

Earnings Estimate for Elementary and Secondary School Teachers

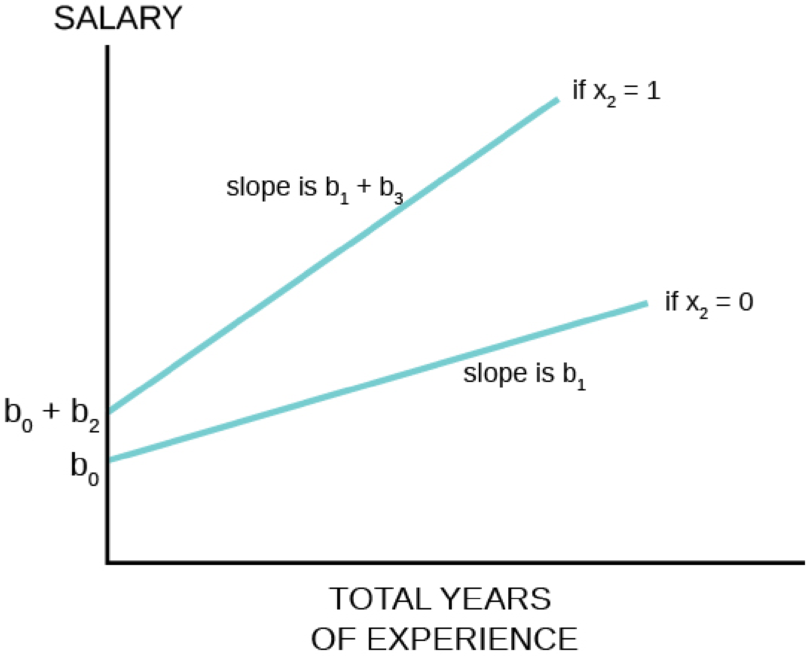
The coefficients for all the independent variables are significantly different from zero as indicated by the standard errors. Dividing the standard errors of each coefficient results in a t-value greater than 1.96 which is the required level for 95% significance. The binary variable, our dummy variable of interest in this analysis, is gender where male is given a value of 1 and female given a value of 0. The coefficient is significantly different from zero with a dramatic t-statistic of 47 standard deviations. We thus cannot accept the null hypothesis that the coefficient is equal to zero. Therefore we conclude that there is a premium paid male teachers of $632 after holding constant experience, education and the wealth of the school district in which the teacher is employed. It is important to note that these data are from some time ago and the $632 represents a six percent salary premium at that time. A graph of this example of dummy variables is presented below.



In two dimensions, salary is the dependent variable on the vertical axis and total years of experience was chosen for the continuous independent variable on horizontal axis. Any of the other independent variables could have been chosen to illustrate the effect of the dummy variable. The relationship between total years of experience has a slope of $52.32 per year of experience and the estimated line has an intercept of $4,269 if the gender variable is equal to zero, for female. If the gender variable is equal to 1, for male, the coefficient for the gender variable is added to the intercept and thus the relationship between total years of experience and salary is shifted upward parallel as indicated on the graph. Also marked on the graph are various points for reference. A female school teacher with 10 years of experience receives a salary of $4,792 on the basis of her experience only, but this is still $109 less than a male teacher with zero years of experience.

A more complex interaction between a dummy variable and the dependent variable can also be estimated. It may be that the dummy variable has more than a simple shift effect on the dependent variable, but also interacts with one or more of the other continuous independent variables. While not tested in the example above, it could be hypothesized that the impact of gender on salary was not a one-time shift, but impacted the value of additional years of experience on salary also. That is, female school teacher's salaries were discounted at the start, and further did not grow at the same rate from the effect of experience as for male school teachers. This would show up as a different slope for the relationship between total years of experience for males than for females. If this is so then females school teachers would not just start behind their male colleagues (as measured by the shift in the estimated regression line), but would fall further and further behind as time and experienced increased.

The graph below shows how this hypothesis can be tested with the use of dummy variables and an interaction variable.



$$\hat{y} = b_0 + b_2x_2 + b_1x_1 + b_3x_2x_1$$

The estimating equation shows how the slope of X1, the continuous random variable experience, contains two parts, b_1 and b_3. This occurs because of the new variable X_2 X_1, called the interaction variable, was created to allow for an effect on the slope of X_1 from changes in X_2, the binary dummy variable. Note that when the dummy variable, $X_2 = 0$ the interaction variable has a value of 0, but when $X_2 = 1$ the interaction variable has a value of X_1. The coefficient b_3 is an estimate of the difference in the coefficient of X_1 when $X_2 = 1$ compared to when $X_2 = 0$. In the example of teacher's salaries, if there is a premium paid to male teachers that affects the rate of increase in salaries from experience, then the rate at which male teachers' salaries rises would be $b_1 + b_3$ and the rate at which female teachers' salaries rise would be simply b_1. This hypothesis can be tested with the hypothesis:

$$H_0 : \beta_3 = 0 \mid \beta_1 = 0, \beta_2 = 0$$
$$H_a : \beta_3 \neq 0 \mid \beta_1 \neq 0, \beta_2 \neq 0$$

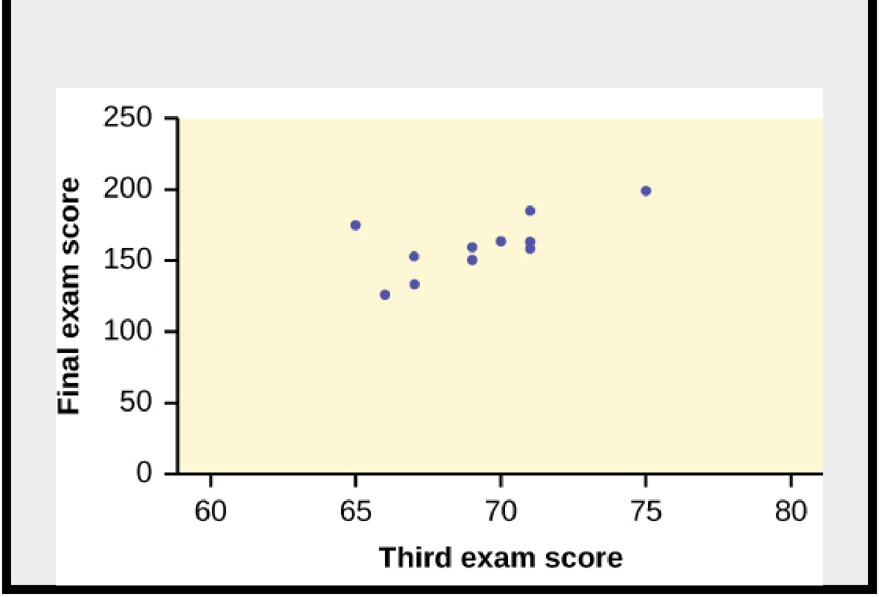
This is a t-test using the test statistic for the parameter β_3. If we cannot accept the null hypothesis that $\beta_3 = 0$ we conclude there is a difference between the rate of increase for the group for whom the value of the binary variable is set to 1, males in this example. This estimating equation can be combined with our earlier one that tested only a parallel shift in the estimated line. The earnings/experience functions in [link] are drawn for this case with a shift in the earnings function and a difference in the slope of the function with respect to total years of experience.

A random sample of 11 statistics students produced the following data, where *x* is the third exam score out of 80, and *y* is the final exam score out of 200. Can you predict the final exam score of a randomly selected student if you know the third exam score?

x (third exam score)	y (final exam score)
65	175
67	133
71	185
71	163
66	126
75	198
67	153
70	163
71	159
69	151
69	159

Table showing the scores on the final exam based on scores from the third exam.

Scatter plot showing the scores on the final exam based on scores from the third exam.



Suppose that you have at your disposal the information below for each of 30 drivers. Propose a model (including a very brief indication of symbols used to represent independent variables) to explain how miles per gallon vary from driver to driver on the basis of the factors measured.

Information:

1. miles driven per day
2. weight of car
3. number of cylinders in car
4. average speed
5. miles per gallon
6. number of passengers

Yj = b0 + b1·X1 + b2·X2 + b3·X3 + b4·X4 + b5·X6 + ej

Consider a sample least squares regression analysis between a dependent variable (Y) and an independent variable (X). A sample correlation coefficient of −1 (minus one) tells us that

1. there is no relationship between Y and X in the sample
2. there is no relationship between Y and X in the population
3. there is a perfect negative relationship between Y and X in the population
4. there is a perfect negative relationship between Y and X in the sample.

d. there is a perfect negative relationship between Y and X in the sample.

In correlational analysis, when the points scatter widely about the regression line, this means that the correlation is

1. negative.
2. low.
3. heterogeneous.
4. between two measures that are unreliable.

b. low

Chapter Review

It is hoped that this discussion of regression analysis has demonstrated the tremendous potential value it has as a tool for testing models and helping to better understand the world around us. The regression model has its limitations, especially the requirement that the underlying relationship be approximately linear. To the extent that the true relationship is nonlinear it may be approximated with a linear relationship or nonlinear forms of transformations that can be estimated with linear techniques. Double logarithmic transformation of the data will provide an easy way to test this particular shape of the relationship. A reasonably good quadratic form (the shape of the total cost curve from Microeconomics Principles) can be generated by the equation:

$$Y = a + b1X + b2X2$$

where the values of X are simply squared and put into the equation as a separate variable.

There is much more in the way of econometric "tricks" that can bypass some of the more troublesome assumptions of the general regression model. This statistical technique is so valuable that further study would provide any student significant, statistically significant, dividends.

Glossary

Residual or “error”

the value calculated from subtracting $y0 - \hat{y}0 = e0$. The absolute value of a residual measures the vertical distance between the actual value of *y* and the estimated value of *y* that appears on the best-fit line.

Sum of Squared Errors (SSE)

the calculated value from adding up all the squared residual terms. The hope is that this value is very small when creating a model.

R2 – Coefficient of Determination

This is a number between 0 and 1 that represents the percentage variation of the dependent variable that can be explained by the variation in the independent variable. Sometimes calculated by the equation $R2 = SSRSST$ where SSR is the “Sum of Squares Regression” and SST is the “Sum of Squares Total.” The appropriate coefficient of determination to be reported should always be adjusted for degrees of freedom first.

Interpretation of Regression Coefficients: Elasticity and Logarithmic Transformation

As we have seen, the coefficient of an equation estimated using OLS regression analysis provides an estimate of the slope of a straight line that is assumed be the relationship between the dependent variable and at least one independent variable. From the calculus, the slope of the line is the first derivative and tells us the magnitude of the impact of a one unit change in the X variable upon the value of the Y variable measured in the units of the Y variable. As we saw in the case of dummy variables, this can show up as a parallel shift in the estimated line or even a change in the slope of the line through an interactive variable. Here we wish to explore the concept of elasticity and how we can use a regression analysis to estimate the various elasticities in which economists have an interest.

The concept of elasticity is borrowed from engineering and physics where it is used to measure a material's responsiveness to a force, typically a physical force such as a stretching/pulling force. It is from here that we get the term an “elastic” band. In economics, the force in question is some market force such as a change in price or income. Elasticity is measured as a percentage change/response in both engineering applications and in economics. The value of measuring in percentage terms is that the units of measurement do not play a role in the value of the measurement and thus allows direct comparison between elasticities. As an example, if the price of gasoline increased say 50 cents from an initial price of $3.00 and generated a decline in monthly consumption for a consumer from 50 gallons to 48 gallons we calculate the elasticity to be 0.25. The price elasticity is the percentage change in quantity resulting from some percentage change in price. A 16 percent increase in price has generated only a 4 percent decrease in demand: 16% price change → 4% quantity change or .04/.16 = .25. This is called an inelastic demand meaning a small response to the price change. This comes about because there are few if any real substitutes for gasoline; perhaps public transportation, a bicycle or walking. Technically, of course, the percentage change in demand from a price increase is a decline in demand thus price elasticity is a negative number. The common convention, however, is to talk about elasticity as the absolute value of the number. Some goods have many substitutes: pears for apples for plums, for grapes, etc. etc. The elasticity for such goods is larger than one and are called elastic in demand. Here a small percentage change in price will induce a large percentage change in quantity demanded. The consumer will easily shift the demand to the close substitute.

While this discussion has been about price changes, any of the independent variables in a demand equation will have an associated elasticity. Thus, there is an income elasticity that measures the sensitivity of demand to changes in income: not much for the demand for food, but very sensitive for yachts. If the demand equation contains a term for substitute goods, say candy bars in a demand equation for cookies, then the responsiveness of demand for cookies from changes in prices of candy bars can be measured. This is called the cross-price elasticity of demand and to an extent can be thought of as brand loyalty from a marketing view. How responsive is the demand for Coca-Cola to changes in the price of Pepsi?

Now imagine the demand for a product that is very expensive. Again, the measure of elasticity is in percentage terms thus the elasticity can be directly compared to that for gasoline: an elasticity of 0.25 for gasoline conveys the same information as an elasticity of 0.25 for $25,000 car. Both goods are considered by the consumer to have few substitutes and thus have inelastic demand curves, elasticities less than one.

The mathematical formulae for various elasticities are:
Price elasticity:$\eta p = (\%\Delta Q)(\%\Delta P)$

Where η is the Greek small case letter eta used to designate elasticity. Δ is read as “change”.
Income elasticity:$\eta Y = (\%\Delta Q)(\%\Delta Y)$

Where Y is used as the symbol for income.
Cross-Price elasticity:$\eta p1 = (\%\Delta Q1)(\%\Delta P2)$

Where P2 is the price of the substitute good.

Examining closer the price elasticity we can write the formula as:
$\eta p = (\%\Delta Q)(\%\Delta P) = dQdP(PQ) = b(PQ)$

Where b is the estimated coefficient for price in the OLS regression.

The first form of the equation demonstrates the principle that elasticities are measured in percentage terms. Of course, the ordinary least squares coefficients provide an estimate of the impact of a unit change in the independent variable, X, on the dependent variable measured in units of Y. These coefficients are not elasticities, however, and are shown in the second way of writing the formula for elasticity as (dQdP), the derivative of the estimated demand function which is simply the slope of the regression line. Multiplying the slope times PQ provides an elasticity measured in percentage terms.

Along a straight-line demand curve the percentage change, thus elasticity, changes continuously as the scale changes, while the slope, the estimated regression coefficient, remains constant. Going back to the demand for gasoline. A change in price from \$3.00 to \$3.50 was a 16 percent increase in price. If the beginning price were $5.00 then the same 50¢ increase would be only a 10 percent increase generating a different elasticity. Every straight-line demand curve has a range of elasticities starting at the top left, high prices, with large elasticity numbers, elastic demand, and decreasing as one goes down the demand curve, inelastic demand.

In order to provide a meaningful estimate of the elasticity of demand the convention is to estimate the elasticity at the point of means. Remember that all OLS regression lines will go through the point of means. At this point is the greatest weight of the data used to estimate the coefficient. The formula to estimate an elasticity when an OLS demand curve has been estimated becomes:

$\eta p = b(P - Q -)$

Where P– and Q– are the mean values of these data used to estimate b, the price coefficient.

The same method can be used to estimate the other elasticities for the demand function by using the appropriate mean values of the other variables; income and price of substitute goods for example.

Logarithmic Transformation of the Data

Ordinary least squares estimates typically assume that the population relationship among the variables is linear thus of the form presented in The Regression Equation. In this form the interpretation of the coefficients is as discussed above; quite simply the coefficient provides an estimate of the impact of a one **unit** change in X on Y measured in **units** of Y. It does not matter just where along the line one wishes to make the measurement because it is a straight line with a constant slope thus constant estimated level of impact per unit change. It may be, however, that the analyst wishes to estimate not the simple unit measured impact on the Y variable, but the magnitude of the percentage impact on Y of a one unit change in the X variable. Such a case might be how a **unit change** in experience, say one year, effects not the absolute amount of a worker's wage, but the **percentage impact** on the worker's wage. Alternatively, it may be that the question asked is the unit measured impact on Y of a specific percentage increase in X. An example may be "by how many dollars will sales increase if the firm spends X percent more on advertising?" The third possibility is the case of elasticity discussed above. Here we are interested in the percentage impact on quantity demanded for a given percentage change in price, or income or perhaps the price of a substitute good. All three of these cases can be estimated by transforming the data to logarithms before running the regression. The resulting coefficients will then provide a percentage change measurement of the relevant variable.

To summarize, there are four cases:

1. Unit ΔX → Unit ΔY (Standard OLS case)
2. Unit ΔX → %ΔY
3. %ΔX → Unit ΔY
4. %ΔX → %ΔY (elasticity case)

Case 1: The ordinary least squares case begins with the linear model developed above:

$$Y=a+bX$$

where the coefficient of the independent variable $b=dYdX$ is the slope of a straight line and thus measures the impact of a unit change in X on Y measured in units of Y.

Case 2: The underlying estimated equation is:

$$\log(Y)=a+bX$$

The equation is estimated by converting the Y values to logarithms and using OLS techniques to estimate the coefficient of the X variable, b. This is called a semi-log estimation. Again, differentiating both sides of the equation allows us to develop the interpretation of the X coefficient b:

$$d(\log Y)=bdX$$

$$dYY=bdX$$

Multiply by 100 to covert to percentages and rearranging terms gives:

$$100b=\%\Delta Y\text{Unit }\Delta X$$

100b is thus the percentage change in Y resulting from a unit change in X.

Case 3: In this case the question is “what is the unit change in Y resulting from a percentage change in X?” What is the dollar loss in revenues of a five percent increase in price or what is the total dollar cost impact of a five percent increase in labor costs? The estimated equation for this case would be:
$Y = a + B\log(X)$

Here the calculus differential of the estimated equation is:
$dY = bd(\log X)$
$dY = bdXX$

Divide by 100 to get percentage and rearranging terms gives:
$b100 = dY100dXX = \text{Unit } \Delta Y\%\Delta X$

Therefore, b100 is the increase in Y measured in units from a one percent increase in X.

Case 4: This is the elasticity case where both the dependent and independent variables are converted to logs before the OLS estimation. This is known as the log-log case or double log case, and provides us with direct estimates of the elasticities of the independent variables. The estimated equation is:
$\log Y = a + b\log X$

Differentiating we have:

$d(logY)=bd(logX)$
$d(logX)=b1XdX$

thus:
$1Y\ dY=b\ 1X\ dX$ OR $dYY\ =b\ dXX$ OR $b=\ dYdX(XY)$

and $b=\%\Delta Y\%\Delta X$ our definition of elasticity. We conclude that we can directly estimate the elasticity of a variable through double log transformation of the data. The estimated coefficient is the elasticity. It is common to use double log transformation of all variables in the estimation of demand functions to get estimates of all the various elasticities of the demand curve.

In a linear regression, why do we need to be concerned with the range of the independent (X) variable?

The precision of the estimate of the Y variable depends on the range of the independent (X) variable explored. If we explore a very small range of the X variable, we won't be able to make much use of the regression. Also, extrapolation is not recommended.

Suppose one collected the following information where X is diameter of tree trunk and Y is tree height.

X	Y	
4	8	
2	4	
8	18	
6	22	
10	30	
6	8	

Regression equation: $y\hat{}i = -3.6 + 3.1\cdot Xi$

What is your estimate of the average height of all trees having a trunk diameter of 7 inches?

$y\hat{} = -3.6 + (3.1\cdot 7) = 18.1$

The manufacturers of a chemical used in flea collars claim that under standard test conditions each additional unit of the chemical will bring about a reduction of 5 fleas (i.e. where $Xj =$ amount of chemical and $YJ = B0 + B1\cdot XJ + EJ$, H0: $B1 = -5$

Suppose that a test has been conducted and results from a computer include:

Intercept $= 60$

Slope $= -4$

Standard error of the regression coefficient $=$ 1.0

Degrees of Freedom for Error $= 2000$

95% Confidence Interval for the slope -2.04, -5.96

Is this evidence consistent with the claim that the number of fleas is reduced at a rate of 5 fleas per unit chemical?

Most simply, since -5 is included in the confidence interval for the slope, we can conclude that the evidence is consistent with the claim at the 95% confidence level.

Using a t test:

H0: $B1 = -5$

HA: $B1 \neq -5$

$\text{tcalculated} = -5 - (-4) 1 = -1$

$\text{tcritical} = -1.96$

Since tcalc $<$ tcrit we retain the null hypothesis that $\text{B1} = -5$.

Predicting with a Regression Equation

One important value of an estimated regression equation is its ability to predict the effects on Y of a change in one or more values of the independent variables. The value of this is obvious. Careful policy cannot be made without estimates of the effects that may result. Indeed, it is the desire for particular results that drive the formation of most policy. Regression models can be, and have been, invaluable aids in forming such policies.

The Gauss-Markov theorem assures us that the point estimate of the impact on the dependent variable derived by putting in the equation the hypothetical values of the independent variables one wishes to simulate will result in an estimate of the dependent variable which is minimum variance and unbiased. That is to say that from this equation comes the best unbiased point estimate of y given the values of x.

$$\hat{y} = b_0 + b_, X_{1i} + \cdots + b_k X_{ki}$$

Remember that point estimates do not carry a particular level of probability, or level of confidence, because points have no “width” above which there is an area to measure. This was why we developed confidence intervals for the mean and proportion earlier. The same concern arises here also. There are actually two different approaches to the issue of developing estimates of changes in the independent variable, or variables, on the dependent variable. The first approach wishes to measure the **expected mean** value of y from a specific change in the value of x: this specific value implies the expected value. Here the question is: what is the **mean** impact on y that would result from multiple hypothetical experiments on y at this specific value of x. Remember that there is a variance around the estimated parameter of x and thus each experiment will result in a bit of a different estimate of the predicted value of y.

The second approach to estimate the effect of a specific value of x on y treats the event as a single experiment: you choose x and multiply it times the coefficient and that provides a single estimate of y. Because this approach acts as if there were a single experiment the variance that exists in the parameter estimate is larger than the variance associated with the expected value approach.

The conclusion is that we have two different ways to predict the effect of values of the independent variable(s) on the dependent variable and thus we have two different intervals. Both are correct answers to the question being asked, but there are two different questions. To avoid confusion, the first case where we are asking for the **expected value** of the mean of the estimated y, is called a **confidence interval** as we have named this concept before. The second case, where we are asking for the estimate of the impact on the dependent variable y of a single experiment using a value of x, is called the **prediction interval**. The test statistics for these two interval measures within which the estimated value of y will fall are:

Confidence Interval for Expected Value of Mean Value of y for $x = x_p$

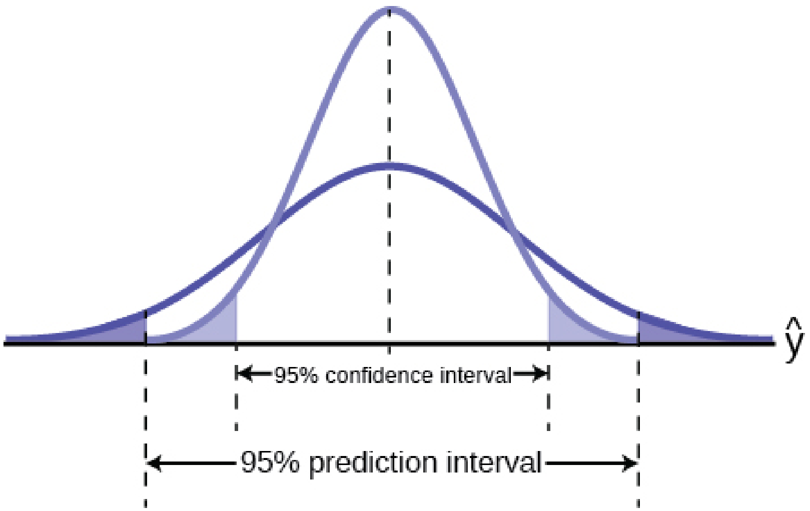
ŷ = ± t α 2 s e (1 n + (x p - x –) 2 sx)

Prediction Interval for an Individual y for $x = x_p$

ŷ = ± t α 2 s e (1 + 1 n + (x p - x –) 2 sx)

Where s_e is the standard deviation of the error term and s_x is the standard deviation of the x variable.

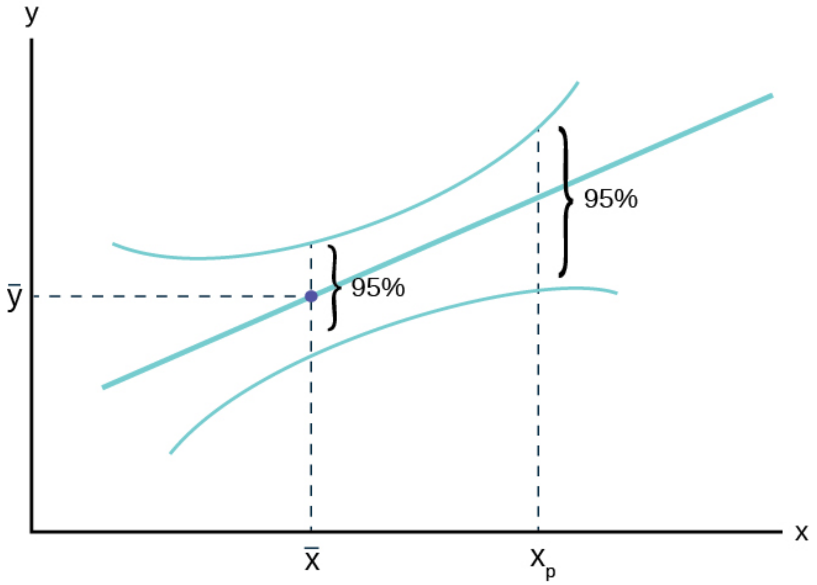
The mathematical computations of these two test statistics are complex. Various computer regression software packages provide programs within the regression functions to provide answers to inquires of estimated predicted values of y given various values chosen for the x variable(s). It is important to know just which interval is being tested in the computer package because the difference in the size of the standard deviations will change the size of the interval estimated. This is shown in [link]. Prediction and confidence intervals for regression equation; 95% confidence level.



[link] shows visually the difference the standard deviation makes in the size of the estimated intervals. The confidence interval, measuring the expected value of the dependent variable, is smaller than the prediction interval for the same level of confidence. The expected value method assumes that the experiment is conducted multiple times rather than just once as in the other method. The logic here is similar, although not identical, to that discussed when developing the relationship between the sample size and the confidence interval using the Central Limit Theorem. There, as the number of experiments increased, the distribution narrowed and the confidence interval became tighter around the expected value of the mean.

It is also important to note that the intervals around a point estimate are highly dependent upon the range of data used to estimate the equation regardless of which approach is being used for prediction. Remember that all regression equations go through the point of means, that is, the mean value of y and the mean values of all independent variables in the equation. As the value of x chosen to estimate the associated value of y is further from the point of means the width of the estimated interval around the point estimate increases. Choosing values of x beyond the range of the data used to estimate the equation possess even greater danger of creating estimates with little use; very large intervals, and risk of error. [link] shows this relationship.

Confidence interval for an individual value of x, Xp, at 95% level of confidence



[link] demonstrates the concern for the quality of the estimated interval whether it is a prediction

interval or a confidence interval. As the value chosen to predict y, X_p in the graph, is further from the central weight of the data, X–, we see the interval expand in width even while holding constant the level of confidence. This shows that the precision of any estimate will diminish as one tries to predict beyond the largest weight of the data and most certainly will degrade rapidly for predictions beyond the range of the data. Unfortunately, this is just where most predictions are desired. They can be made, but the width of the confidence interval may be so large as to render the prediction useless. Only actual calculation and the particular application can determine this, however.

Recall the third exam/final exam example .

We found the equation of the best-fit line for the final exam grade as a function of the grade on the third-exam. We can now use the least-squares regression line for prediction. Assume the coefficient for X was determined to be significantly different from zero.

Suppose you want to estimate, or predict, the mean final exam score of statistics students who received 73 on the third exam. The exam scores **(*x*-values)** range from 65 to 75. Since 73 is between the *x*-values 65 and 75, we feel comfortable to substitute $x = 73$ into the equation. Then:

$$\hat{y} = -173.51 + 4.83(73) = 179.08$$

We predict that statistics students who earn a grade of 73 on the third exam will earn a grade of 179.08 on the final exam, on average.

a. What would you predict the final exam score to be for a student who scored a 66 on the third exam?

a. 145.27

b. What would you predict the final exam score to be for a student who scored a 90 on the third exam?

b. The *x* values in the data are between 65 and 75. Ninety is outside of the domain of the observed *x* values in the data (independent variable), so you cannot reliably predict the final exam score for this student. (Even though it is possible to enter 90 into the equation for *x* and calculate a corresponding *y* value, the *y* value that you get will have a confidence interval that may not be meaningful.)

To understand really how unreliable the prediction can be outside of the observed *x* values observed in the data, make the substitution $x = 90$ into the equation.

$\hat{y} = -173.51 + 4.83(90) = 261.19$

The final-exam score is predicted to be 261.19. The largest the final-exam score can be is 200.

True or False? If False, correct it: Suppose you are performing a simple linear regression of Y on X and you test the hypothesis that the slope β is zero against a two-sided alternative. You have $n = 25$ observations and your computed test (t) statistic is 2.6. Then your P-value is given by $.01 < P < .02$, which gives borderline significance (i.e. you would reject H0 at $\alpha = .02$ but fail to reject H0 at $\alpha = .01$).

True.

$t_{(\text{critical, df} = 23\text{, two-tailed, } \alpha = .02)} = \pm 2.5$

$t_{\text{critical, df} = 23\text{, two-tailed, } \alpha = .01} = \pm 2.8$

An economist is interested in the possible influence of "Miracle Wheat" on the average yield of wheat in a district. To do so he fits a linear regression of average yield per year against year after introduction of "Miracle Wheat" for a ten year period.

The fitted trend line is

$\hat{y}j = 80 + 1.5 \cdot Xj$

(Yj: Average yield in *j* year after introduction)

(Xj: *j* year after introduction).

1. What is the estimated average yield for the fourth year after introduction?
2. Do you want to use this trend line to estimate yield for, say, 20 years after introduction? Why? What would your estimate be?

---

1. $80 + 1.5 \cdot 4 = 86$
2. No. Most business statisticians would not want to extrapolate that far. If someone did, the estimate would be 110, but some other factors probably come into play with 20 years.

An interpretation of $r = 0.5$ is that the following part of the Y-variation is associated with which variation in X:

1. most
2. half
3. very little
4. one quarter
5. none of these

---

d. one quarter

Which of the following values of *r* indicates the most accurate prediction of one variable from another?

1. $r = 1.18$
2. $r = -.77$
3. $r = .68$

---

b. $r = -.77$

How to Use Microsoft Excel® for Regression Analysis

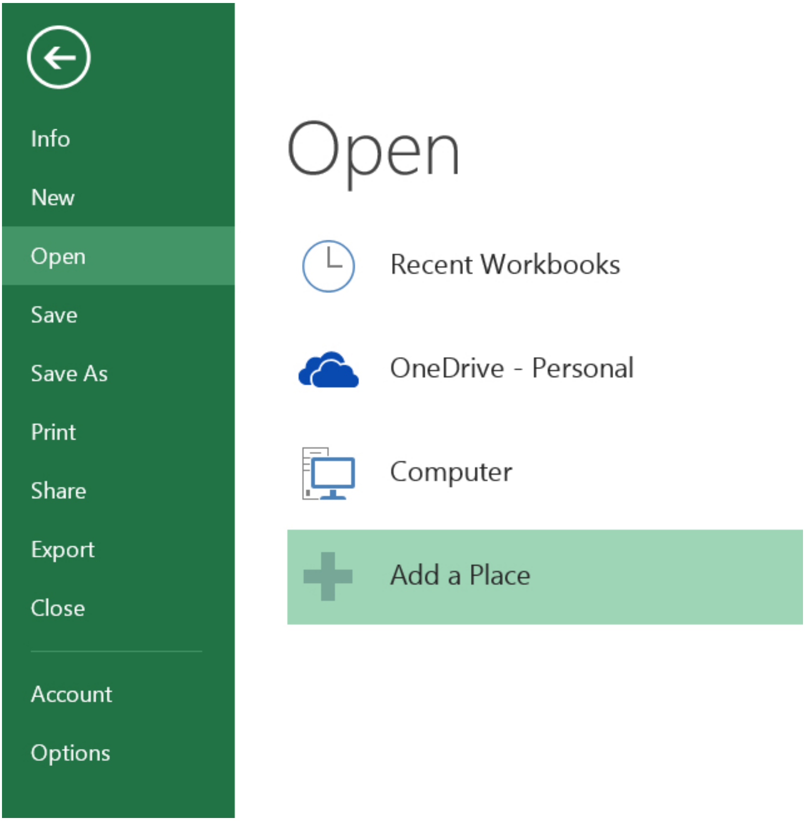
This section of this chapter is here in recognition that what we are now asking requires much more than a quick calculation of a ratio or a square root. Indeed, the use of regression analysis was almost non- existent before the middle of the last century and did not really become a widely used tool until perhaps the late 1960's and early 1970's. Even then the computational ability of even the largest IBM machines is laughable by today's standards. In the early days programs were developed by the researchers and shared. There was no market for something called "software" and certainly nothing called "apps", an entrant into the market only a few years old.

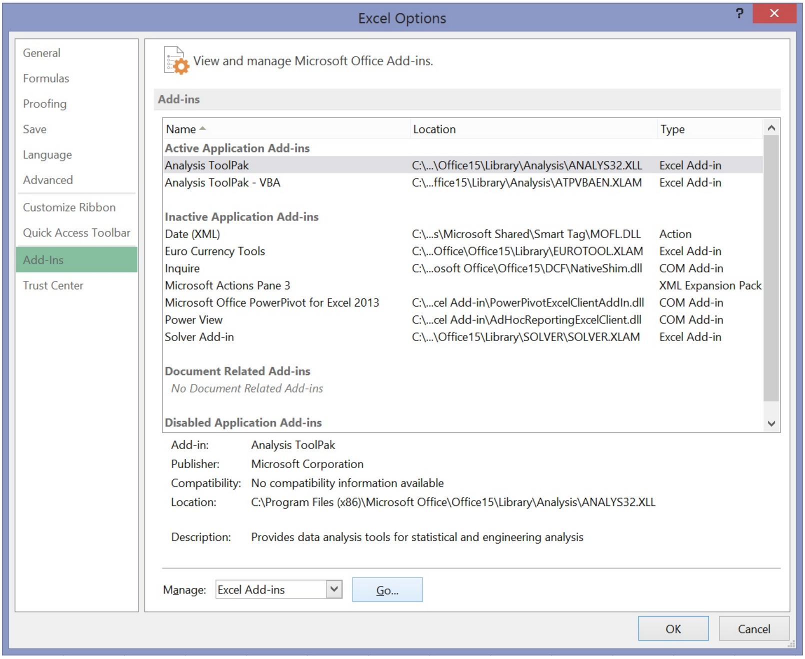
With the advent of the personal computer and the explosion of a vital software market we have a number of regression and statistical analysis packages to choose from. Each has their merits. We have chosen Microsoft Excel because of the wide-spread availability both on college campuses and in the post-college market place. Stata is an alternative and has features that will be important for more advanced econometrics study if you choose to follow this path. Even more advanced packages exist, but typically require the analyst to do some significant amount of programing to conduct their analysis. The goal of this section is to demonstrate how to use Excel to run a regression and then to do so with an example of a simple version of a demand curve.

The first step to doing a regression using Excel is to load the program into your computer. If you have Excel you have the Analysis ToolPak although you may not have it activated. The program calls upon a significant amount of space so is not loaded automatically.

To activate the Analysis ToolPak follow these steps:

Click “File” > “Options” > “Add-ins” to bring up a menu of the add-in “ToolPaks”. Select “Analysis ToolPak” and click “GO” next to “Manage: excel add-ins” near the bottom of the window. This will open a new window where you click “Analysis ToolPak” (make sure there is a green check mark in the box) and then click “OK”. Now there should be an Analysis tab under the data menu. These steps are presented in the following screen shots.


Info
New
Open
Save
Save As
Print
Share
Export
Close
Account
Options
Open
Recent Workbooks
OneDrive - Personal
Computer
Add a Place



View and manage Microsoft Office Add-ins.

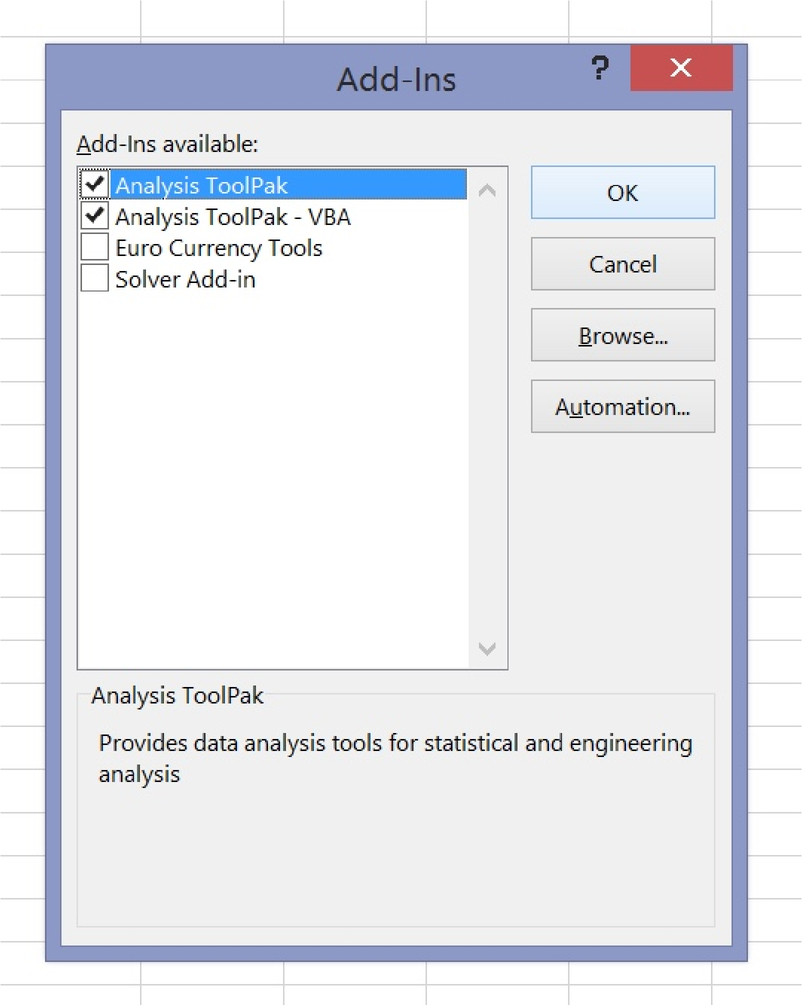
Add-ins

Name	Location	Type
Active Application Add-ins		
Analysis ToolPak	C:\...\Office15\Library\Analysis\ANALYS32.XLL	Excel Add-in
Analysis ToolPak - VBA	C:\...ffice15\Library\Analysis\ATPVBAEN.XLAM	Excel Add-in
Inactive Application Add-ins		
Date (XML)	C:\...s\Microsoft Shared\Smart Tag\MOFL.DLL	Action
Euro Currency Tools	C:\...Office\Office15\Library\EUROTOOL.XLAM	Excel Add-in
Inquire	C:\...osoft Office\Office15\DCF\NativeShim.dll	COM Add-in
Microsoft Actions Pane 3		XML Expansion Pack
Microsoft Office PowerPivot for Excel 2013	C:\...cel Add-in\PowerPivotExcelClientAddIn.dll	COM Add-in
Power View	C:\...cel Add-in\AdHocReportingExcelClient.dll	COM Add-in
Solver Add-in	C:\...\Office15\Library\SOLVER\SOLVER.XLAM	Excel Add-in
Document Related Add-ins		
No Document Related Add-ins		
Disabled Application Add-ins		

Add-in: Analysis ToolPak
Publisher: Microsoft Corporation
Compatibility: No compatibility information available
Location: C:\Program Files (x86)\Microsoft Office\Office15\Library\Analysis\ANALYS32.XLL

Description: Provides data analysis tools for statistical and engineering analysis

Manage: Excel Add-ins



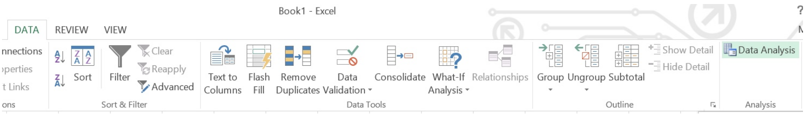
Add-Ins

Add-Ins available:

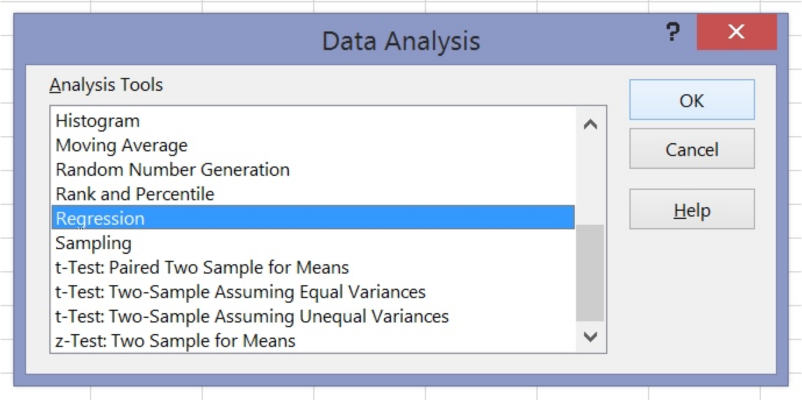
- ☑ Analysis ToolPak
- ☑ Analysis ToolPak - VBA
- ☐ Euro Currency Tools
- ☐ Solver Add-in

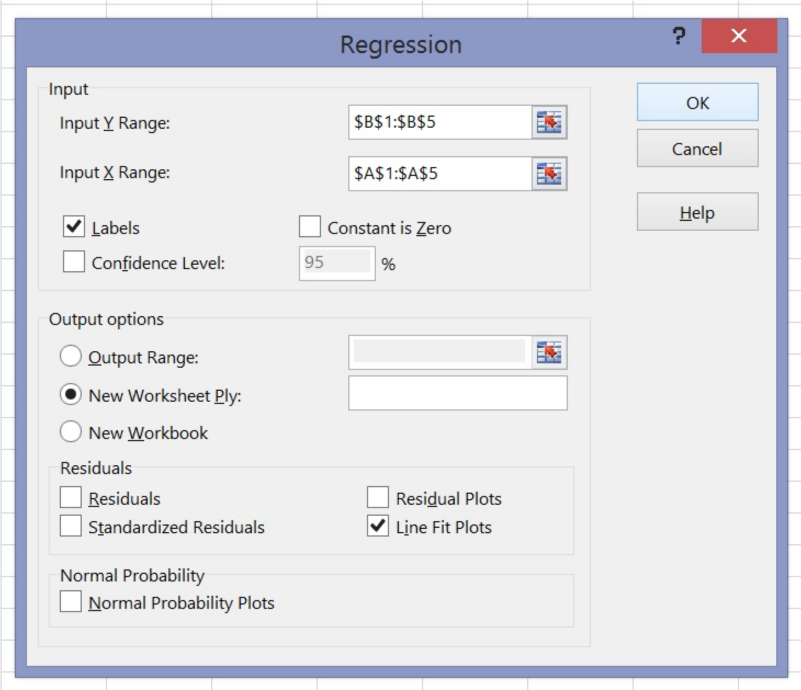
Analysis ToolPak

Provides data analysis tools for statistical and engineering analysis



Click “Data” then “Data Analysis” and then click “Regression” and “OK”. Congratulations, you have made it to the regression window. The window asks for your inputs. Clicking the box next to the Y and X ranges will allow you to use the click and drag feature of Excel to select your input ranges. Excel has one odd quirk and that is the click and drop feature requires that the independent variables, the X variables, are all together, meaning that they form a single matrix. If your data are set up with the Y variable between two columns of X variables Excel will not allow you to use click and drag. As an example, say Column A and Column C are independent variables and Column B is the Y variable, the dependent variable. Excel will not allow you to click and drop the data ranges. The solution is to move the column with the Y variable to column A and then you can click and drag. The same problem arises again if you want to run the regression with only some of the X variables. You will need to set up the matrix so all the X variables you wish to regress are in a tightly formed matrix. These steps are presented in the following scene shots.


Data Analysis
?
×
Analysis Tools
Histogram
Moving Average
Random Number Generation
Rank and Percentile
Regression
Sampling
t-Test: Paired Two Sample for Means
t-Test: Two-Sample Assuming Equal Variances
t-Test: Two-Sample Assuming Unequal Variances
z-Test: Two Sample for Means
OK
Cancel
Help



Input

Input Y Range: $B$1:$B$5

Input X Range: $A$1:$A$5

☑ Labels   □ Constant is Zero

□ Confidence Level: 95 %

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Residuals

□ Residuals   □ Residual Plots

□ Standardized Residuals   ☑ Line Fit Plots

Normal Probability

□ Normal Probability Plots

Once you have selected the data for your regression analysis and told Excel which one is the dependent variable (Y) and which ones are the independent valuables (X‘s), you have several choices as to the parameters and how the output will be displayed. Refer to screen shot [link] under “Input” section. If you check the “labels” box the program will place the entry in the first column of each variable as its name in the output. You can enter an actual name, such as price or income in a demand analysis, in row one of the Excel spreadsheet for each variable and it will be displayed in the output.

The level of significance can also be set by the analyst. This will not change the calculated t statistic, called t stat, but will alter the p value for the calculated t statistic. It will also alter the boundaries of the confidence intervals for the coefficients. A 95 percent confidence interval is always presented, but with a change in this you will also get other levels of confidence for the intervals.

Excel also will allow you to suppress the intercept. This forces the regression program to minimize the residual sum of squares under the condition that the estimated line must go through the origin. This is done in cases where there is no meaning in the model at some value other than zero, zero for the start of the line. An example is an economic production function that is a relationship between the number of units of an input, say hours of labor, and output. There is no meaning of positive output with zero workers.

Once the data are entered and the choices are made click OK and the results will be sent to a separate new worksheet by default. The output from Excel is presented in a way typical of other regression package programs. The first block of information gives the overall statistics of the regression: Multiple R, R Squared, and the R squared adjusted for degrees of freedom, which is the one you want to report. You also get the Standard error (of the estimate) and the number of observations in the regression.

The second block of information is titled ANOVA which stands for Analysis of Variance. Our interest in this section is the column marked F. This is the calculated F statistics for the null hypothesis that all of the coefficients are equal to zero verse the alternative that at least one of the coefficients are not equal to zero. This hypothesis test was presented in 13.4 under "How Good is the Equation?" The next column gives the p value for this test under the title "Significance F". If the p value is less than say 0.05 (the calculated F statistic is in the tail) we can say with 90 % confidence that we cannot accept the null hypotheses that all the coefficients are equal to zero. This is a good thing: it means that at least one of the coefficients is significantly different from zero thus do have an effect on the value of Y.

The last block of information contains the hypothesis tests for the individual coefficient. The estimated coefficients, the intercept and the slopes, are first listed and then each standard error (of the estimated coefficient) followed by the t stat (calculated student's t statistic for the null hypothesis that the coefficient is equal to zero). We compare the t stat and the critical value of the student's t, dependent on the degrees of freedom, and determine if we have enough evidence to reject the null that the variable has no effect on Y. Remember that we have set up the null hypothesis as the status quo and our claim that we know what caused the Y to change is in the alternative hypothesis. We want to reject the status quo and substitute our version of the world, the alternative hypothesis. The next column contains the p values for this hypothesis test followed by the estimated upper and lower bound of the confidence interval of the estimated slope parameter for various levels of confidence set by us at the beginning.

Estimating the Demand for Roses

Here is an example of using the Excel program to run a regression for a particular specific case: estimating the demand for roses. We are trying to estimate a demand curve, which from economic theory we expect certain variables affect how much of a good we buy. The relationship between the price of a good and the quantity demanded is the demand curve. Beyond that we have the demand function that includes other relevant variables: a person’s income, the price of substitute goods, and perhaps other variables such as season of the year or the price of complimentary goods. Quantity demanded will be our Y variable, and Price of roses, Price of carnations and Income will be our independent variables, the X variables.

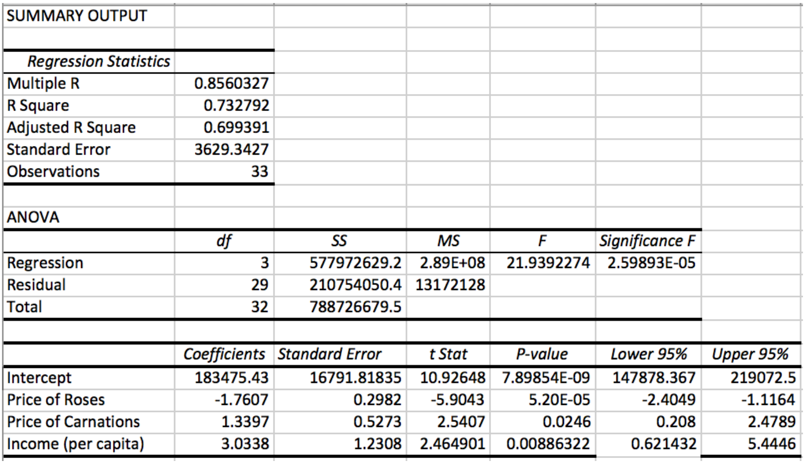
For all of these variables theory tells us the expected relationship. For the price of the good in question, roses, theory predicts an inverse relationship, the negatively sloped demand curve. Theory also predicts the relationship between the quantity demanded of one good, here roses, and the price of a substitute, carnations in this example. Theory predicts that this should be a positive or direct relationship; as the price of the substitute falls we substitute away from roses to the cheaper substitute, carnations. A reduction in the price of the substitute generates a reduction in demand for the good being analyzed, roses here. Reduction generates reduction is a positive relationship. For normal goods, theory also predicts a positive relationship; as our incomes rise we buy more of the good, roses. We expect these results because that is what is predicted by a hundred years of economic theory and research. Essentially we are testing these century-old hypotheses. The data gathered was determined by the model that is being tested. This should always be the case. One is not doing inferential statistics by throwing a mountain of data into a computer and asking the machine for a theory. Theory first, test follows.

These data here are national average prices and income is the nation's per capita personal income. Quantity demanded is total national annual sales of roses. These are annual time series data; we are tracking the rose market for the United States from 1984-2017, 33 observations.

Because of the quirky way Excel requires how the data are entered into the regression package it is best to have the independent variables, price of roses, price of carnations and income next to each other on the spreadsheet. Once your data are entered into the spreadsheet it is always good to look at the data. Examine the range, the means and the standard deviations. Use your understanding of descriptive statistics from the very first part of this course. In large data sets you will not be able to "scan" the data. The Analysis ToolPac makes it easy to get the range, mean, standard deviations and other parameters of the distributions. You can also quickly get the correlations among the variables. Examine for outliers. Review the history. Did something happen? Was here a labor strike, change in import fees, something that makes these observations unusual? Do not take the data without question. There may have been a typo somewhere, who knows without review.

Go to the regression window, enter the data and select 95% confidence level and click "OK". You can include the labels in the input range if you have put a title at the top of each column, but be sure to click the "labels" box on the main regression page if you do.

The regression output should show up automatically on a new worksheet.



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.8560327
R Square	0.732792
Adjusted R Square	0.699391
Standard Error	3629.3427
Observations	33

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	577972629.2	2.89E+08	21.9392274	2.59893E-05
Residual	29	210754050.4	13172128		
Total	32	788726679.5			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	183475.43	16791.81835	10.92648	7.89854E-09	147878.367	219072.5
Price of Roses	-1.7607	0.2982	-5.9043	5.20E-05	-2.4049	-1.1164
Price of Carnations	1.3397	0.5273	2.5407	0.0246	0.208	2.4789
Income (per capita)	3.0338	1.2308	2.464901	0.00886322	0.621432	5.4446

The first results presented is the R-Square, a measure of the strength of the correlation between Y and X1, X2, and X3 taken as a group. Our R-square here of 0.699, adjusted for degrees of freedom, means that 70% of the variation in Y, demand for roses, can be explained by variations in X1, X2, and X3, Price of roses, Price of carnations and Income. There is no statistical test to determine the "significance" of an R2. Of course a higher R2 is preferred, but it is really the significance of the coefficients that will determine the value of the theory being tested and which will become part of any policy discussion if they are demonstrated to be significantly different form zero.

Looking at the third panel of output we can write the equation as:

Y = b0 + b1X1 + b2X2 + b3X3 + e

where b0 is the intercept, b1 is the estimated coefficient on price of roses, and b2 is the estimated coefficient on price of carnations, b3 is the estimated effect of income and e is the error term. The equation is written in Roman letters indicating that these are the estimated values and not the population parameters, β's.

Our estimated equation is:
Quantity of roses sold = 183,475 − 1.76Price of roses + 1.33Price of carnations + 3.03Income

We first observe that the signs of the coefficients are as expected from theory. The demand curve is downward sloping with the negative sign for the price of roses. Further the signs of both the price of carnations and income coefficients are positive as would be expected from economic theory.

Interpreting the coefficients can tell us the magnitude of the impact of a change in each variable on the demand for roses. It is the ability to do this which makes regression analysis such a valuable tool. The estimated coefficients tell us that an increase the price of roses by one dollar will lead to a 1.76 reduction in the number roses purchased. The price of carnations seems to play an important role in the demand for roses as we see that increasing the price of carnations by one dollar would increase the demand for roses by 1.33 units as consumers would substitute away from the now more expensive carnations. Similarly, increasing per capita income by one dollar will lead to a 3.03 unit increase in roses purchased.

These results are in line with the predictions of economics theory with respect to all three variables included in this estimate of the demand for roses. It is important to have a theory first that predicts the significance or at least the direction of the coefficients. Without a theory to test, this research tool is not much more helpful than the correlation coefficients we learned about earlier.

We cannot stop there, however. We need to first check whether our coefficients are statistically significant from zero. We set up a hypothesis of:
$H0 : \beta 1 = 0$
$Ha : \beta 1 \neq 0$

for all three coefficients in the regression. Recall from earlier that we will not be able to definitively say that our estimated b_1 is the actual real population of β_1, but rather only that with $(1-\alpha)\%$ level of confidence that we cannot reject the null hypothesis that our estimated β_1 is significantly different from zero. The analyst is making a claim that the price of roses causes an impact on quantity demanded. Indeed, that each of the included variables has an impact on the quantity of roses demanded. The claim is therefore in the alternative hypotheses. It will take a very large probability, 0.95 in this case, to overthrow the null hypothesis, the status quo, that $\beta = 0$. In all regression hypothesis tests the claim is in the alternative and the claim is that the theory has found a variable that has a significant impact on the Y variable.

The test statistic for this hypothesis follows the familiar standardizing formula which counts the number of standard deviations, t, that the estimated value of the parameter, b_1, is away from the hypothesized value, β_0, which is zero in this case:
$tc = b1 - \beta 0 \ Sb1$

The computer calculates this test statistic and presents it as “t stat”. You can find this value to the right of the standard error of the coefficient estimate. The standard error of the coefficient for b_1 is S_{b1} in the formula. To reach a conclusion we compare this test statistic with the critical value of the student’s t at degrees of freedom n-3-1 =29, and alpha = 0.025 (5% significance level for a two-tailed test). Our t stat for b_1 is approximately 5.90 which is greater than 1.96 (the critical value we looked up in the t-table), so we cannot accept our null hypotheses of no effect. We conclude that Price has a significant effect because the calculated t value is in the tail. We conduct the same test for b_2 and b_3. For each variable, we find that we cannot accept the null hypothesis of no relationship because the calculated t-statistics are in the tail for each case, that is, greater than the critical value. All variables in this regression have been determined to have a significant effect on the demand for roses.

These tests tell us whether or not an individual coefficient is significantly different from zero, but does not address the overall quality of the model. We have seen that the R squared adjusted for degrees of freedom indicates this model with these three variables explains 70% of the variation in quantity of roses demanded. We can also conduct a second test of the model taken as a whole. This is the F test presented in section 13.4 of this chapter. Because this is a multiple regression (more than one X), we use the F-test to determine if our coefficients collectively affect Y. The hypothesis is:
$H_0 : \beta_1 = \beta_2 = \ldots = \beta_i = 0$
H_a : "at least one of the β_i is not equal to 0"

Under the ANOVA section of the output we find the calculated F statistic for this hypotheses. For this example the F statistic is 21.9. Again, comparing the calculated F statistic with the critical value given our desired level of significance and the degrees of freedom will allow us to reach a conclusion.

The best way to reach a conclusion for this statistical test is to use the p-value comparison rule. The p-value is the area in the tail, given the calculated F statistic. In essence the computer is finding the F value in the table for us and calculating the p-value. In the Summary Output under “significance F” is this probability. For this example, it is calculated to be 2.6 x 10^{-5}, or 2.6 then moving the decimal five places to the left. (.000026) This is an almost infinitesimal level of probability and is certainly less than our alpha level of .05 for a 5 percent level of significance.

By not being able to accept the null hypotheses we conclude that this specification of this model has validity because at least one of the estimated coefficients is significantly different from zero. Since F-calculated is greater than F-critical, we cannot accept H_0, meaning that X_1, X_2 and X_3 *together* has a significant effect on Y.

The development of computing machinery and the software useful for academic and business research has made it possible to answer questions that just a few years ago we could not even formulate. Data is available in electronic format and can be moved into place for analysis in ways and at speeds that were unimaginable a decade ago. The sheer magnitude of data sets that can today be used for research and analysis gives us a higher quality of results than in days past. Even with only an Excel spreadsheet we can conduct very high level research. This section gives you the tools to conduct some of this very interesting research with the only limit being your imagination.

A computer program for multiple regression has been used to fit y^j = b0 + b1·X1j + b2·X2j + b3·X3j.

Part of the computer output includes:

i	bi	Sbi
0	8	1.6
1	2.2	.24
2	.72	.32
3	0.005	0.002

1. Calculation of confidence interval for b2 consists of _____ ± (a student's t value) (_____)
2. The confidence level for this interval is reflected in the value used for _____.
3. The degrees of freedom available for estimating the variance are directly concerned with the value used for _____

---

1. −.72, .32
2. the t value
3. the t value

An investigator has used a multiple regression program on 20 data points to obtain a regression equation with 3 variables. Part of the computer output is:

Variable	**Coefficient**	**Standard Error of bi**
1	0.45	0.21
2	0.80	0.10
3	3.10	0.86

1. 0.80 is an estimate of _______.
2. 0.10 is an estimate of _______.
3. Assuming the responses satisfy the normality assumption, we can be 95% confident that the value of β2 is in the interval,_____ ± [t.025 · _____], where t.025 is the critical value of the student's t distribution with ___ degrees of freedom.

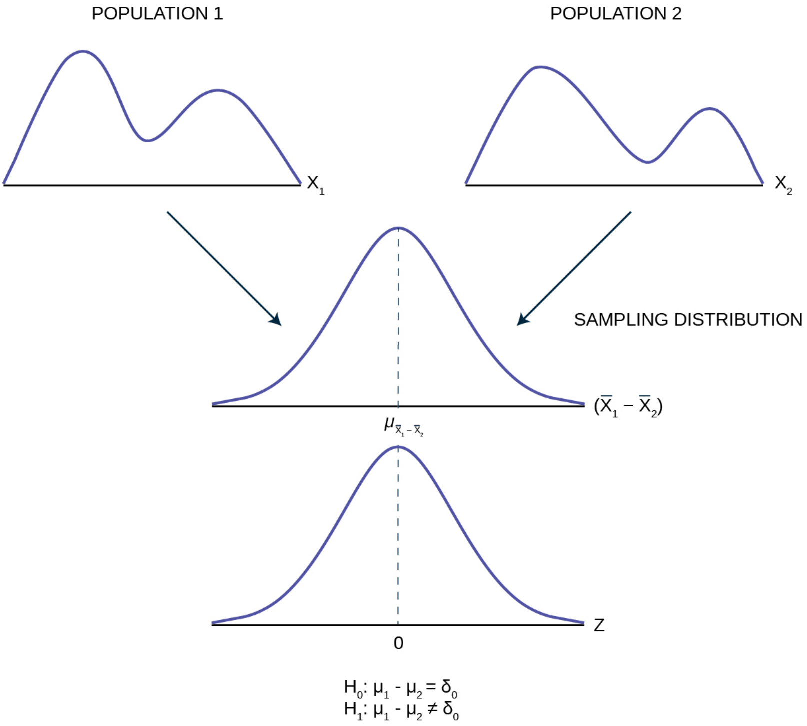
---

1. The population value for β2, the change that occurs in Y with a unit change in X2, when the other variables are held constant.
2. The population value for the standard error of the distribution of estimates of β2.
3. .8, .1, $16 = 20 - 4$.

Test for Differences in Means Large and Small Samples

The comparison of two independent population means is very common and provides a way to test the hypothesis that the two groups differ from each other. Is the night shift less productive than the day shift, are the rates of return from fixed asset investments different from those from common stock investments, and so on? Any observed difference between two sample means depends on both the means themselves and the standard deviations. Very different means can occur by chance if there is great variation among the individual samples. The test statistic will have to account for this fact. So, just as when working with single population tests, we need to consider our sample size. If both samples are sufficiently large (n1 ≥ 30 and n2 ≥ 30), we can assume the associated sampling distribution will be approximately normal. In such cases we will use a z-based approach to any tests we conduct. However, if one or both samples are small (i.e. at least one of the sample contains fewer than 30 observations), we cannot assume the associated sampling distribution will be normally shaped unless both populations are also normally shaped. In these cases we will use a t-based approach.

Now we are interested in whether or not two samples have the same mean. Our question has not changed: Do these two samples come from the same population distribution? To approach this problem we create a new random variable. We recognize that we have two sample means, one from each set of data, and thus we have two random variables coming from two unknown distributions. To solve the problem we create a new random variable, the difference between the sample means. This new random variable also has a distribution and, again, the Central Limit Theorem tells us that this new distribution is normally distributed, regardless of the underlying distributions of the original data. A graph may help to understand this concept.


POPULATION 1
POPULATION 2
X_1
X_2
SAMPLING DISTRIBUTION
$(\overline{X}_1 - \overline{X}_2)$
$\mu_{\overline{X}_1 - \overline{X}_2}$
Z
0
$H_0: \mu_1 - \mu_2 = \delta_0$
$H_1: \mu_1 - \mu_2 \neq \delta_0$

Pictured are two distributions of data, X1 and X2, with unknown means and standard deviations. The second panel shows the sampling distribution of the newly created random variable (X-1-X-2). This distribution is the theoretical distribution of many many sample means from population 1 minus sample means from population 2. The Central Limit Theorem tells us that this theoretical sampling distribution of differences in sample means is normally distributed, regardless of the distribution of the actual population data shown in the top panel. Because the sampling distribution is normally distributed, we can develop a standardizing formula and calculate probabilities from the standard normal distribution in the bottom panel, the Z distribution. We have seen this same analysis before in Chapter 7 Figure 7.2 .

The Central Limit Theorem, as before, provides us with the standard deviation of the sampling distribution, and further, that the expected value of the mean of the distribution of differences in sample means is equal to the differences in the population means. Mathematically this can be stated:

$$E (\mu x\text{-}1 - \mu x\text{-}2) = \mu 1 - \mu 2$$

If we know the population standard deviations, we do not need to estimate them. Therefore, for the hypothesis test, we use the two known population standard deviations to calculate the estimated standard deviation of the sampling distribution, or

standard error, of **the difference in sample means**, X ¯ 1 – X ¯ 2 .

The standard error is:

(σ 1) 2 n 1 + (σ 2) 2 n 2

Recalling how a Z score was calculated when working with one population confidence intervals and hypothesis tests, the new formula is closely related but incorporates two sample means and two population standard deviations.**The test statistic (*Z*-score) is calculated as follows:**

Zc = (x ¯ 1 – x ¯ 2)–δ0 (σ 1) 2 n 1 + (σ 2) 2 n 2

If we do not know the population standard deviations, we will simply substitute the sample standard deviations into the above formula. This does, of course, add some error to our estimate of the standard error. As long as both samples are sufficiently large, this added error will be small. The revised standard error formula will become as follows:

The standard error is:

(s 1) 2 n 1 + (s 2) 2 n 2

The test statistic (*t*-score) is calculated as follows:

tc = (x ¯ 1 – x ¯ 2)–δ0 (s 1) 2 n 1 + (s 2) 2 n 2

Degrees of freedom for equal population varainces

Because of the added error when working with small sample sizes, we must also account for degrees of freedom when using a t-based approach. When both samples have been drawn from populations that have equal variances (standard deviations) then the degrees of freedom can be calculated as n1 + n2 - 2. (Later we will learn how to test for the assumption of equal variances.)

Degrees of freedom for unequal population variances

However, if we cannot assume both populations have equal variances, the number of **degrees of freedom (*df*)** requires a somewhat complicated calculation. The formula for *df* when working with unequal variances is as follows:

Degrees of freedom for unequal population variances

df= ((s 1) 2 n 1 + (s 2) 2 n 2) 2 (1 n 1 –1) ((s 1) 2 n 1) 2 +(1 n 2 –1) ((s 2) 2 n 2) 2

NOTE: This *df* calculation does not often result in a whole number.To be conservative, always round this calculation up to the nearest whole number.

The format of the sampling distribution, differences in sample means, specifies that the format of the null and alternative hypothesis is:

$H0 : \mu1 - \mu2 = \delta0$

$Ha : \mu1 - \mu2 \neq \delta0$

where δ_0 is the hypothesized difference between the two means. If the question is simply “is there any difference between the means?” then $\delta_0 = 0$ and the null and alternative hypotheses becomes:

$H0 : \mu 1 = \mu 2$
$Ha : \mu 1 \neq \mu 2$

An example of when δ_0 might not be zero is when the comparison of the two groups requires a specific difference for the decision to be meaningful. Imagine that you are making a capital investment. You are considering changing from your current model machine to another. You measure the productivity of your machines by the speed they produce the product. It may be that a contender to replace the old model is faster in terms of product throughput, but is also more expensive. The second machine may also have more maintenance costs, setup costs, etc. The null hypothesis would be set up so that the new machine would have to be better than the old one by enough to cover these extra costs in terms of speed and cost of production. This form of the null and alternative hypothesis shows how valuable this particular hypothesis test can be. For most of our work we will be testing simple hypotheses asking if there is any difference between the two distribution means.

Independent groups and small samples

The Kona Iki Corporation produces coconut milk.

They take coconuts and extract the milk inside by drilling a hole and pouring the milk into a vat for processing. They have both a day shift (called the B shift) and a night shift (called the G shift) to do this part of the process. They would like to know if the day shift and the night shift are equally efficient in processing the coconuts. A study is done sampling 9 shifts of the G shift and 16 shifts of the B shift. The results of the number of hours required to process 100 pounds of coconuts is presented in [link]. A study is done and data are collected, resulting in the data in [link]. We will assume at this point that the population variances are unequal.

	Sample Size	Average Number of Hours to Process 100 Pounds of Coconuts	Sample Standard Deviation
G Shift	9	2	0.866
B Shift	16	3.2	1.00

Is there a difference in the mean amount of time for each shift to process 100 pounds of coconuts? Test at the 5% level of significance.

The population standard deviations are not known and cannot be assumed to equal each other. Moreover, the samples are small. Therefore we will use a t-based approach using the unequal variances procedure. . Let *g* be the subscript for the G Shift and *b* be the subscript for the B Shift. Then, μ_g is the population mean for G Shift and μ_b is the population mean for B Shift. This is a test of two **independent groups**, two population **means**.

Random variable: $\bar{X}g - \bar{X}b$ = difference in the sample mean amount of time between the G Shift and the B Shift takes to process the coconuts.

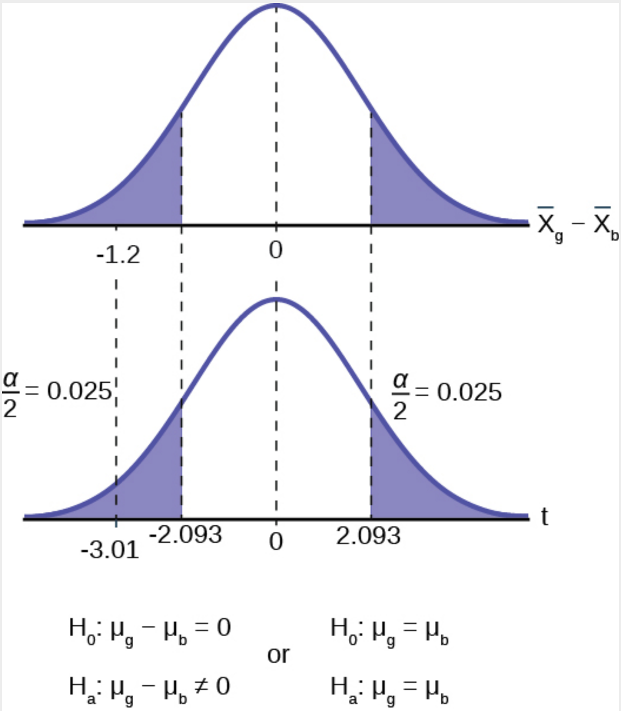
H_0: $\mu_g = \mu_b$ H_0: $\mu_g - \mu_b = 0$

H_a: $\mu_g \neq \mu_b$ H_a: $\mu_g - \mu_b \neq 0$

The words **"the same"** tell you H_0 has an "=". Since there are no other words to indicate H_a, is either faster or slower. This is a two tailed test.

Distribution for the test: Use t_{df} where *df* is calculated using the *df* formula for independent groups, two population means above. Using a calculator, *df* is approximately 18.8462, which will be rounded up to 19.

Graph:



tc = (X-1 – X-2) – δ0 S12n1 + S22n2 = -3.14

We next find the critical value from Excel using the T.INV function and the degrees of freedom from above. The critical value, 2.093, is found using α/2 (0.025) and 19 degrees of freedom. (The convention is to round up the degrees of freedom to make the conclusion more conservative.) Next we calculate the test statistic using the formula above.

Make a decision: Since the calculated t-value is in the tail (that is, it is smaller than the critical value), we must reject the null hypothesis that there is no difference between the two groups. The means are different. Alternatively, we can make a decision by comparing the p-value to alpha. Using the T-DIST function on Excel results in a p-value of 0.01. Since this p-value is smaller than alpha, we can reject the null hypothesis.

The graph has included the sampling distribution of the differences in the sample means to show how the t-distribution aligns with the sampling distribution data. We see in the top panel that the calculated difference in the two means is -1.2 and the bottom panel shows that this is 3.01 standard deviations from the mean. Typically we do not need to show the sampling distribution graph and can rely on the graph of the test statistic, the t-distribution in this case, to reach our conclusion.

Conclusion: At the 5% level of significance, the sample data show there is sufficient evidence to conclude that the mean number of hours that the G Shift takes to process 100 pounds of coconuts is different from the B Shift (mean number of hours for the B Shift is greater than the mean number of hours for the G Shift).

A study is done to determine if Company A retains its workers longer than Company B. It is believed that Company A has a higher retention than Company B. The study finds that in a sample of 31 workers at Company A their average time with the company is four years with a standard deviation of 1.5 years. A sample of 39 workers at Company B finds that the average time with the company was 3.5 years with a standard deviation of 1 year. Test this proposition at the 1% level of significance.

a. Is this a test of two means or two proportions?

a. two means because time is a continuous random variable.

b. Are the populations standard deviations known or unknown?

b. unknown

c. Which distribution do you use to perform the test?

c. A Z-based test can be used since both samples are greater than 30.

d. What is the random variable?

d. X ¯ A - X ¯ B

e. What are the null and alternate hypotheses?

e.

- $H\ o : \mu\ A \leq \mu\ B$
- $H\ a : \mu\ A > \mu\ B$

f. Is this test right-, left-, or two-tailed?

f. right one-tailed test

g. What is the value of the test statistic?

Zc = (X-1 − X-2) − δ0 S12n1 + S22n2 = 1.6

h. Can you accept/reject the null hypothesis?

h. Cannot reject the null hypothesis that there is no difference between the two groups. The test statistic of 1.6 is not in the tail. The critical value of the Z distribution is 2.33. Alternatively, the p-value is 0.0553, which is greater than alpha (0.01), which again supports the conclusion that we cannot reject the null hypothesis.

i. **Conclusion:**

i. At the 1% level of significance, from the sample data, there is not sufficient evidence to conclude that the retention of workers at Company A is longer than Company B, on average.

An interesting research question is the effect, if any, that different types of teaching formats have on the grade outcomes of students. To investigate this issue one sample of students' grades was taken from a hybrid class and another sample taken from a standard lecture format class. Both classes were for the same subject. The mean course grade in percent for the 35 hybrid students is 74 with a standard deviation of 16. The mean grades of the 40 students form the standard lecture class was 76 percent with a standard deviation of 9. Test at 5% to see if there is any significant difference in the population mean grades between standard lecture course and hybrid class.

We begin by noting that we have two groups, students from a hybrid class and students from a standard lecture format class. We also note that the random variable, what we are interested in, is students' grades, a continuous random variable. We could have asked the research question in a different way and had a binary random variable. For example, we could have studied the percentage of students with a failing grade, or with an A grade. Both of these would be binary and thus a test of proportions and not a test of means as is the case here. Finally, there is no presumption as to which format might lead to higher grades so the hypothesis is stated as a two-tailed test.

$H_0: \mu_1 = \mu_2$

Ha: $\mu 1 \neq \mu 2$

As would virtually always be the case, we do not know the population variances of the two distributions and thus our test statistic is:
$Zc=(\bar{x}1-\bar{x}2)-\delta 0s2n1+s2n2=(74-76)-016235+9$
-0.65

Conclusion: The critical Z score is -1.96. Since the test statistic (-0.65) is NOT less than the critical value, we cannot reject the null at α = 5%. Therefore, evidence does not exist to prove that the grades in hybrid and standard classes differ. Alternatively, the p-value is 0.5128, which is greater than alpha (0.05), supporting the above conclusion.

References

Data from Graduating Engineer + Computer Careers. Available online at http://www.graduatingengineer.com

Data from *Microsoft Bookshelf*.

Data from the United States Senate website, available online at www.Senate.gov (accessed June 17, 2013).

“List of current United States Senators by Age.” Wikipedia. Available online at http://en.wikipedia.org/wiki/List_of_current_United_States_Senators_by_age (accessed June 17, 2013).

“Sectoring by Industry Groups.” Nasdaq. Available online at http://www.nasdaq.com/markets/barchart-sectors.aspx?page=sectors&base=industry (accessed June 17, 2013).

“Strip Clubs: Where Prostitution and Trafficking Happen.” Prostitution Research and Education, 2013. Available online at www.prostitutionresearch.com/ProsViolPosttrauStress.html (accessed June 17, 2013).

“World Series History.” Baseball-Almanac, 2013. Available online at http://www.baseball-almanac.com/ws/wsmenu.shtml (accessed June 17, 2013).

Chapter Review

Two population means from independent samples where the population standard deviations are not known

- Random Variable: $X – 1 - X – 2 =$ the difference of the sampling means
- Distribution: Student's *t*-distribution with degrees of freedom (variances not pooled)

Formula Review

Standard error: $SE = (s 1) 2 n 1 + (s 2) 2 n 2$

The test statistic when working with two large samples is: $Zc = (x ¯ 1 – x ¯ 2)–δ0 (σ 1) 2 n 1 + (σ 2) 2 n 2$

If the population standard deviations are unknown, we can still compute the Z-based test statistic by substituting the samples standard deviations into the above formula.

The test statistic when working with at least one small sample is (*t*-score): $tc = (x ¯ 1 – x ¯ 2)–δ0 (s 1) 2 n 1 + (s 2) 2 n 2$

Degrees of freedom:

df= ((s 1) 2 n 1 + (s 2) 2 n 2) 2 (1 n 1 −1) ((s 1) 2 n 1) 2 +(1 n 2 −1) ((s 2) 2 n 2) 2

where:

s1 and s2 are the sample standard deviations, and n1 and n2 are the sample sizes.

x ¯ 1 and x ¯ 2 are the sample means.

Use the following information to answer the next 15 exercises: Indicate if the hypothesis test is for

1. independent group means, population standard deviations, and/or variances known
2. independent group means, population standard deviations, and/or variances unknown
3. matched or paired samples
4. single mean
5. two proportions
6. single proportion

It is believed that 70% of males pass their drivers test in the first attempt, while 65% of females pass the test in the first attempt. Of interest is whether the proportions are in fact equal.

---

two proportions

A new laundry detergent is tested on consumers. Of interest is the proportion of consumers who prefer the new brand over the leading competitor. A study is done to test this.

A new windshield treatment claims to repel water more effectively. Ten windshields are tested by simulating rain without the new treatment. The same windshields are then treated, and the experiment is run again. A hypothesis test is conducted.

---

matched or paired samples

The known standard deviation in salary for all mid-level professionals in the financial industry is $11,000. Company A and Company B are in the financial industry. Suppose samples are taken of mid-level professionals from Company A and from Company B. The sample mean salary for mid-level professionals in Company A is $80,000. The sample mean salary for mid-level professionals in Company B is $96,000. Company A and Company B management want to know if their mid-level professionals are paid differently, on average.

The average worker in Germany gets eight weeks of paid vacation.

---

single mean

According to a television commercial, 80% of dentists agree that Ultrafresh toothpaste is the best on the market.

It is believed that the average grade on an English essay in a particular school system for females is higher than for males. A random sample of 31 females had a mean score of 82 with a standard deviation of three, and a random sample of 25 males had a mean score of 76 with a standard deviation of four.

---

independent group means, population standard deviations and/or variances unknown

The league mean batting average is 0.280 with a known standard deviation of 0.06. The Rattlers and the Vikings belong to the league. The mean batting average for a sample of eight Rattlers is 0.210, and the mean batting average for a sample of eight Vikings is 0.260. There are 24 players on the Rattlers and 19 players on the

Vikings. Are the batting averages of the Rattlers and Vikings statistically different?

In a random sample of 100 forests in the United States, 56 were coniferous or contained conifers. In a random sample of 80 forests in Mexico, 40 were coniferous or contained conifers. Is the proportion of conifers in the United States statistically more than the proportion of conifers in Mexico?

---

two proportions

A new medicine is said to help improve sleep. Eight subjects are picked at random and given the medicine. The means hours slept for each person were recorded before starting the medication and after.

It is thought that teenagers sleep more than adults on average. A study is done to verify this. A sample of 16 teenagers has a mean of 8.9 hours slept and a standard deviation of 1.2. A sample of 12 adults has a mean of 6.9 hours slept and a standard deviation of 0.6.

---

independent group means, population standard deviations and/or variances unknown

Varsity athletes practice five times a week, on average.

A sample of 12 in-state graduate school programs at school A has a mean tuition of $64,000 with a standard deviation of $8,000. At school B, a sample of 16 in-state graduate programs has a mean of $80,000 with a standard deviation of $6,000. On average, are the mean tuitions different?

---

independent group means, population standard deviations and/or variances unknown

A new WiFi range booster is being offered to consumers. A researcher tests the native range of 12 different routers under the same conditions. The ranges are recorded. Then the researcher uses the new WiFi range booster and records the new ranges. Does the new WiFi range booster do a better job?

A high school principal claims that 30% of student athletes drive themselves to school, while 4% of non-athletes drive themselves to school. In a sample of 20 student athletes, 45% drive themselves to school. In a sample of 35 non-athlete students, 6% drive themselves to school. Is the percent of student athletes who drive themselves to school more than the percent of nonathletes?

---

two proportions

Use the following information to answer the next three exercises: A study is done to determine which of two soft drinks has more sugar. There are 13 cans of Beverage A in a sample and six cans of Beverage B. The mean amount of sugar in Beverage A is 36 grams with a standard deviation of 0.6 grams. The mean amount of sugar in Beverage B is 38 grams with a standard deviation of 0.8 grams. The researchers believe that Beverage B has more sugar than Beverage A, on average. Both populations have normal distributions.

Are standard deviations known or unknown?

What is the random variable?

The random variable is the difference between the mean amounts of sugar in the two soft drinks.

Is this a one-tailed or two-tailed test?

Use the following information to answer the next 12 exercises: The U.S. Center for Disease Control reports that the mean life expectancy was 47.6 years for whites born in 1900 and 33.0 years for nonwhites. Suppose that you randomly survey death records for people born in 1900 in a certain county. Of the 124 whites, the mean life span was 45.3 years with a standard deviation of 12.7 years. Of the 82 nonwhites, the mean life span was 34.1 years with a standard deviation of 15.6 years. Conduct a hypothesis test to see if the mean life spans in the county were the same for whites and nonwhites.

Is this a test of means or proportions?

---

means

State the null and alternative hypotheses.

1. H_0: _______
2. H_a: _______

Is this a right-tailed, left-tailed, or two-tailed test?

two-tailed

In symbols, what is the random variable of interest for this test?

In words, define the random variable of interest for this test.

the difference between the mean life spans of whites and nonwhites

Which distribution (normal or Student's *t*) would you use for this hypothesis test?

Explain why you chose the distribution you did for [link].

This is a comparison of two population means with unknown population standard deviations.

Calculate the test statistic.

Sketch a graph of the situation. Label the horizontal axis. Mark the hypothesized difference and the sample difference. Shade the area corresponding to the *p*-value.

Check student’s solution.

At a pre-conceived $\alpha = 0.05$, what is your:

1. Decision:
2. Reason for the decision:
3. Conclusion (write out in a complete sentence):

---

1. Cannot accept the null hypothesis
2. p-value < 0.05
3. There is not enough evidence at the 5% level of significance to support the claim that life expectancy in the 1900s is different between whites and nonwhites.

Does it appear that the means are the same? Why or why not?

Homework

The mean number of English courses taken in a two–year time period by male and female college students is believed to be about the same. An experiment is conducted and data are collected from 29 males and 16 females. The males took an average of three English courses with a standard deviation of 0.8. The females took an average of four English courses with a standard deviation of 1.0. Are the means statistically the same?

A student at a four-year college claims that mean enrollment at four–year colleges is higher than at two–year colleges in the United States. Two surveys are conducted. Of the 35 two–year colleges surveyed, the mean enrollment was 5,068 with a standard deviation of 4,777. Of the 35 four-year colleges surveyed, the mean enrollment was 5,466 with a standard deviation of 8,191.

Subscripts: 1: two-year colleges; 2: four-year colleges

1. $H0: \mu1 \geq \mu2$
2. $Ha: \mu1 < \mu2$
3. $\bar{X}1 - \bar{X}2$ is the difference between the mean enrollments of the two-year colleges

and the four-year colleges.

4. Student's-*t*
5. test statistic: -0.2480
6. *p*-value: 0.4019
7. Check student's solution.
8.
    1. Alpha: 0.05
    2. Decision: Cannot reject
    3. Reason for Decision: *p*-value > alpha
    4. Conclusion: At the 5% significance level, there is sufficient evidence to conclude that the mean enrollment at four-year colleges is higher than at two-year colleges.

At Rachel's 11th birthday party, eight girls were timed to see how long (in seconds) they could hold their breath in a relaxed position. After a two-minute rest, they timed themselves while jumping. The girls thought that the mean difference between their jumping and relaxed times would be zero. Test their hypothesis.

Relaxed time (seconds)		**Jumping time (seconds)**

26	21	
47	40	
30	28	
22	21	
23	25	
45	43	
37	35	
29	32	

Mean entry-level salaries for college graduates with mechanical engineering degrees and electrical engineering degrees are believed to be approximately the same. A recruiting office thinks that the mean mechanical engineering salary is actually lower than the mean electrical engineering salary. The recruiting office randomly surveys 50 entry level mechanical engineers and 60 entry level electrical engineers. Their mean salaries were $46,100 and $46,700, respectively. Their standard deviations were $3,450 and $4,210, respectively. Conduct a hypothesis test to determine if you agree that the mean entry-level mechanical engineering salary is lower than the mean entry-level electrical engineering salary.

Subscripts: 1: mechanical engineering; 2:

electrical engineering

1. $H_0: \mu_1 \geq \mu_2$
2. $H_a: \mu_1 < \mu_2$
3. $\bar{X}_1 - \bar{X}_2$ is the difference between the mean entry level salaries of mechanical engineers and electrical engineers.
4. t_{108}
5. test statistic: $t = –0.82$
6. *p*-value: 0.2061
7. Check student's solution.
8.
    1. Alpha: 0.05
    2. Decision: Cannot reject the null hypothesis.
    3. Reason for Decision: *p*-value > alpha
    4. Conclusion: At the 5% significance level, there is insufficient evidence to conclude that the mean entry-level salaries of mechanical engineers is lower than that of electrical engineers.

Marketing companies have collected data implying that teenage girls use more ring tones on their cellular phones than teenage boys do. In one particular study of 40 randomly chosen teenage girls and boys (20 of each) with cellular phones, the mean number of ring tones for the girls was 3.2 with a standard deviation of 1.5. The mean for the boys was 1.7 with a standard deviation of 0.8. Conduct a hypothesis test to determine if the means are approximately the same or if the girls' mean is higher than the boys' mean.

Use the information from Appendix C: Data Sets to answer the next four exercises.

Using the data from Lap 1 only, conduct a hypothesis test to determine if the mean time for completing a lap in races is the same as it is in practices.

---

1. $H0{:}\mu 1 = \mu 2$
2. $Ha{:}\mu 1 \neq \mu 2$
3. $\bar{X}1 - \bar{X}2$ is the difference between the mean times for completing a lap in races and in practices.
4. $t_{20.32}$
5. test statistic: –4.70
6. *p*-value: 0.0001
7. Check student's solution.
8.
    1. Alpha: 0.05
    2. Decision: Cannot accept the null hypothesis.
    3. Reason for Decision: *p*-value < alpha
    4. Conclusion: At the 5% significance level, there is sufficient evidence to

conclude that the mean time for completing a lap in races is different from that in practices.

Repeat the test in [link], but use Lap 5 data this time.

Repeat the test in [link], but this time combine the data from Laps 1 and 5.

---

1. $H_0: \mu_1 = \mu_2$
2. $H_a: \mu_1 \neq \mu_2$
3. is the difference between the mean times for completing a lap in races and in practices.
4. $t_{40.94}$
5. test statistic: –5.08
6. *p*-value: zero
7. Check student’s solution.
8.
   1. Alpha: 0.05
   2. Decision: Cannot accept the null hypothesis.
   3. Reason for Decision: *p*-value < alpha
   4. Conclusion: At the 5% significance level, there is sufficient evidence to conclude that the mean time for completing a lap in races is different

from that in practices.

In two to three complete sentences, explain in detail how you might use Terri Vogel's data to answer the following question. "Does Terri Vogel drive faster in races than she does in practices?"

Use the following information to answer the next two exercises. The Eastern and Western Major League Soccer conferences have a new Reserve Division that allows new players to develop their skills. Data for a randomly picked date showed the following annual goals.

Western	**Eastern**
Los Angeles 9	D.C. United 9
FC Dallas 3	Chicago 8
Chivas USA 4	Columbus 7
Real Salt Lake 3	New England 6
Colorado 4	MetroStars 5
San Jose 4	Kansas City 3

Conduct a hypothesis test to answer the next two *exercises.*

The **exact** distribution for the hypothesis test is:

1. the normal distribution
2. the Student's *t*-distribution
3. the uniform distribution
4. the exponential distribution

If the level of significance is 0.05, the conclusion is:

1. There is sufficient evidence to conclude that the **W** Division teams score fewer goals, on average, than the **E** teams
2. There is insufficient evidence to conclude that the **W** Division teams score more goals, on average, than the **E** teams.
3. There is insufficient evidence to conclude that the **W** teams score fewer goals, on average, than the **E** teams score.
4. Unable to determine

---

c

Suppose a statistics instructor believes that there is no significant difference between the mean class scores of statistics day students on

Exam 2 and statistics night students on Exam 2. She takes random samples from each of the populations. The mean and standard deviation for 35 statistics day students were 75.86 and 16.91. The mean and standard deviation for 37 statistics night students were 75.41 and 19.73. The “day” subscript refers to the statistics day students. The “night” subscript refers to the statistics night students. A concluding statement is:

1. There is sufficient evidence to conclude that statistics night students' mean on Exam 2 is better than the statistics day students' mean on Exam 2.
2. There is insufficient evidence to conclude that the statistics day students' mean on Exam 2 is better than the statistics night students' mean on Exam 2.
3. There is insufficient evidence to conclude that there is a significant difference between the means of the statistics day students and night students on Exam 2.
4. There is sufficient evidence to conclude that there is a significant difference between the means of the statistics day students and night students on Exam 2.

Researchers interviewed street prostitutes in Canada and the United States. The mean age of the 100 Canadian prostitutes upon entering prostitution was 18 with a standard deviation of six. The mean age of the 130 United States prostitutes upon entering prostitution was 20 with a standard deviation of eight. Is the mean age of entering prostitution in Canada lower than the mean age in the United States? Test at a 1% significance level.

Test: two independent sample means, population standard deviations unknown.

Random variable: $\bar{X}_1 - \bar{X}_2$

Distribution: H0:µ1=µ2 Ha:µ1<µ2H_0: $\mu_1 = \mu_2$ H_a: $\mu_1 < \mu_2$ The mean age of entering prostitution in Canada is lower than the mean age in the United States.

Graph: left-tailed

p-value : 0.0151

Decision: Cannot reject H_0.

Conclusion: At the 1% level of significance, from the sample data, there is not sufficient evidence to conclude that the mean age of entering prostitution in Canada is lower than the mean age in the United States.

A powder diet is tested on 49 people, and a liquid diet is tested on 36 different people. Of interest is whether the liquid diet yields a higher mean weight loss than the powder diet. The powder diet group had a mean weight loss of 42 pounds with a standard deviation of 12 pounds. The liquid diet group had a mean weight loss of 45 pounds with a standard deviation of 14 pounds.

Suppose a statistics instructor believes that there is no significant difference between the mean class scores of statistics day students on Exam 2 and statistics night students on Exam 2. She takes random samples from each of the populations. The mean and standard deviation for 35 statistics day students were 75.86 and 16.91, respectively. The mean and standard deviation for 37 statistics night students were 75.41 and 19.73. The "day" subscript refers to the statistics day students. The "night" subscript refers to the statistics night students. An appropriate alternative hypothesis for the hypothesis test is:

1. $\mu_{\text{day}} > \mu_{\text{night}}$
2. $\mu_{\text{day}} < \mu_{\text{night}}$
3. $\mu_{\text{day}} = \mu_{\text{night}}$
4. $\mu_{\text{day}} \neq \mu_{\text{night}}$

d

Glossary

Cohen’s *d*

a measure of effect size based on the differences between two means. If *d* is between 0 and 0.2 then the effect is small. If *d* approaches is 0.5, then the effect is medium, and if *d* approaches 0.8, then it is a large effect.

Pooled Variance

a weighted average of two variances that can then be used when calculating standard error.

Test for Differences in Means for Small Samples with Unequal Population Variances

Typically we can never expect to know any of the population parameters, mean, proportion, or standard deviation. When testing hypotheses concerning differences in means we are faced with the difficulty of two unknown variances that play a critical role in the test statistic. We have been substituting the sample variances just as we did when testing hypotheses for a single mean. Moreover, when working with large samples, where both n1 and n2 > 30, we retained a Z-based approach to hypothesis testing (though many statisticians ALWAYS use a t-based approach if the population variances are unknown). When working with smaller samples, however, we use a t-based approach, which compensates for the added error of using small samples taken from populations with unknown variances. There may be situations, however, when we do not know the population variances but can assume that the two populations have the same variance. If this is true, then the pooled sample variance will be smaller than the individual sample variances. This will give more precise estimates and reduce the probability of discarding a good null. The null and alternative hypotheses remain the same, but the test statistic changes to:

$$tc=(\bar{x}1-\bar{x}2)-\delta0Sp2(1n1+1n2)$$

As you can see, the standard error is calculated by using a single, pooled variance of the two samples, rather than by using two separate sample variances. So how do we get this pooled variance? We take the two individual sample variances and take a weighted average of the two of them, as demonstrated below. Keep in mind that if the two sample sizes are the same, then the previous standard error formula and this weighted average formula will give you the same result.

The the pooled variance Sp2 is given by the formula:
$Sp2=(n1-1)s12+(n2-1)s22n1+n2-2$

We will program the above formula into an Excel spreadsheet. That way, you will not have to continually work out the results manually.

A drug trial is attempted using a real drug and a pill made of just sugar. 18 people are given the real drug in hopes of increasing the production of endorphins. The increase in endorphins is found to be on average 8 micrograms per person, and the sample standard deviation is 5.4 micrograms. 11 people are given the sugar pill, and their average endorphin increase is 4 micrograms with a standard deviation of 2.4. From previous research on endorphins it is determined that it can be assumed that the variances within the two samples can be assumed to be the same. Test at 5% to see if the population mean for the real drug had a significantly greater impact on the endorphins than the population mean with the sugar pill.

First we begin by designating one of the two groups Group 1 and the other Group 2. This will be needed to keep track of the null and alternative hypotheses. Let's set Group 1 as those who received the actual new medicine being tested and therefore Group 2 is those who received the sugar pill. We can now set up the null and alternative hypothesis as:

$H_0: \mu_1 \leq \mu_2$
$H_1: \mu_1 > \mu_2$

This is set up as a one-tailed test with the claim in the alternative hypothesis that the medicine will produce more endorphins than the sugar pill. We now calculate the test statistic which requires us to calculate the pooled variance, Sp2 using the formula above.
$tc = (\bar{x}1 - \bar{x}2) - \delta 0Sp2(1n1 + 1n2) = (8 - 4) - 020.4933($

t_α, allows us to compare the test statistic and

the critical value.
$t\alpha = 1.703$ at $df = n1 + n2 - 2 = 18 + 11 - 2 = 27$

The test statistic is clearly in the tail, 2.31 is larger than the critical value of 1.703, and therefore we cannot maintain the null hypothesis. Thus, we conclude that there is significant evidence at the 5% level of significance that the new medicine produces the effect desired. Using the p-value method we get a p-value of 0.01, which is well below our chosen level of alpha (0.05). This supports the conclusion to reject the null hypothesis.

Chapter Review

In situations when we do not know the population variances but assume the variances are the same, the pooled sample variance will be smaller than the individual sample variances.

This will give more precise estimates and reduce the probability of discarding a good null.

Formula Review

$$tc = (\bar{x}1 - \bar{x}2) - \delta0Sp2(1n1 + 1n2)$$

where Sp2 is the pooled variance given by the formula:

$$Sp2 = (n1 - 1)s12 + (n2 - 1)s22n1 + n2 - 2$$

Test for Differences in Population Proportions

When conducting a hypothesis test that compares two independent population proportions, the following characteristics should be present:

1. The two independent samples are simple random samples that are independent.
2. The number of successes is at least five, and the number of failures is at least five, for each of the samples.
3. Growing literature states that the population must be at least ten or 20 times the size of the sample. This keeps each population from being over-sampled and causing incorrect results.

Comparing two proportions, like comparing two means, is common. If two estimated proportions are different, it may be due to a difference in the populations or it may be due to chance. A hypothesis test can help determine if a difference in the estimated proportions reflects a difference in the population proportions.

The difference of two proportions follows an approximate normal distribution. Generally, the null hypothesis states that the two proportions are the same. That is, H_0: $p_A = p_B$. To conduct the test, we use a pooled proportion, p_c.

The pooled proportion is calculated as follows:

$p\ c = x\ A + x\ B\ n\ A + n\ B$

The distribution for the differences is:

$P'_A - P'_B \sim N[0, p_c(1- p_c)(1 n_A + 1 n_B)]$

The test statistic (*z*-score) is:

$z= (p'_A - p'_B)-(p_A - p_B) p_c(1- p_c)(1 n_A + 1 n_B)$

Two types of medication for hives are being tested to determine if there is a **difference in the proportions of adult patient reactions. Twenty** out of a random **sample of 200** adults given medication A still had hives 30 minutes after taking the medication. **Twelve** out of another **random sample of 200 adults** given medication B still had hives 30 minutes after taking the medication. Test at a 1% level of significance.

The problem asks for a difference in proportions, making it a test of two proportions.

Let *A* and *B* be the subscripts for medication A and medication B, respectively. Then p_A and p_B are the desired population proportions.

Random Variable:
$P'_A - P'_B =$ difference in the proportions of

adult patients who did not react after 30 minutes to medication A and to medication B.

$H_0: p_A = p_B$

$p_A – p_B = 0$

$H_a: p_A \neq p_B$

$p_A – p_B \neq 0$

The words "**is a difference**" tell you the test is two-tailed.

Distribution for the test: Since this is a test of two binomial population proportions, the distribution is normal:

p c = x A + x B n A + n B = 20+12 200+200 =0.08   1– p c =0.92

P ′ A – P ′ B ~N[0, (0.08)(0.92)(1 200 + 1 200)]

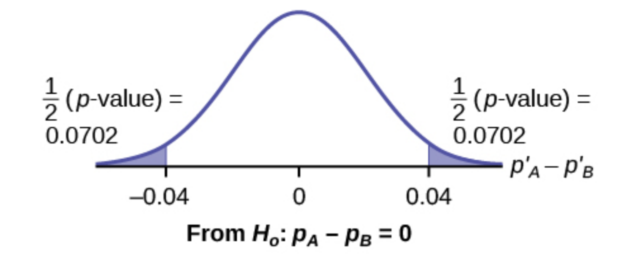
$P'_A – P'_B$ follows an approximate normal distribution.

Calculate the *p*-value using the normal distribution: *p*-value = 0.1404.

Estimated proportion for group A: p ′ A = x A n A = 20 200 =0.1

Estimated proportion for group B: $p\,'\,B = x\,B$ $n\,B = 12\ 200 = 0.06$

Graph:



$P'A - P'B = 0.1 - 0.06 = 0.04.$

Half the *p*-value is below –0.04, and half is above 0.04.

Compare *α* and the *p*-value: $\alpha = 0.01$ and the *p*-value $= 0.1404$. $\alpha < p$-value.

Make a decision: Since $\alpha < p$-value, do not reject *Ho*.

Conclusion: At a 1% level of significance, from the sample data, there is not sufficient evidence to conclude that there is a difference in the proportions of adult patients who did not react after 30 minutes to medication *A* and medication *B*.

Press STAT. Arrow over to TESTS and press 6:2-PropZTest. Arrow down and enter 20 for x1, 200 for n1, 12 for x2, and 200 for n2. Arrow down to p1: and arrow to not equal p2. Press ENTER. Arrow down to Calculate and press ENTER. The *p*-value is $p = 0.1404$ and the test statistic is 1.47. Do the procedure again, but instead of Calculate do Draw.

Try It

Two types of valves are being tested to determine if there is a difference in pressure tolerances. Fifteen out of a random sample of 100 of Valve *A* cracked under 4,500 psi. Six out of a random sample of 100 of Valve *B* cracked under 4,500 psi. Test at a 5% level of significance.

A research study was conducted about gender differences in "sexting." The researcher believed that the proportion of girls involved in "sexting" is less than the proportion of boys involved. The data collected in the spring of 2010 among a random sample of middle and high school students in a large school district in the southern United States is summarized in [link]. Is the proportion of girls sending sexts less than the proportion of boys "sexting?" Test at a 1% level of significance.

	Males	Females
Sent "sexts"	183	156
Total number surveyed	2231	2169

This is a test of two population proportions. Let M and F be the subscripts for males and females. Then p_M and p_F are the desired population proportions.

Random variable:
$p'_F - p'_M =$ difference in the proportions of males and females who sent "sexts."

H_0: $p_F = p_M$ H_0: $p_F - p_M = 0$

H_a: $p_F < p_M$ H_a: $p_F - p_M < 0$

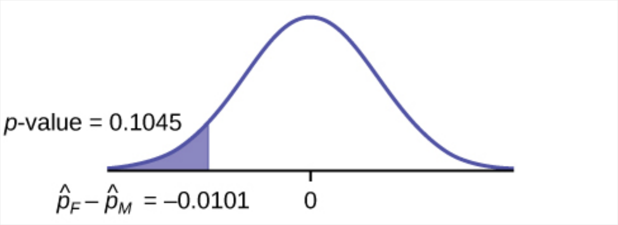
The words "**less than**" tell you the test is left-tailed.

Distribution for the test: Since this is a test of two population proportions, the distribution is normal:

$p_c = \frac{x_F + x_M}{n_F + n_M} = \frac{156+183}{2169+2231} = 0.077$
$1 - p_c = 0.923$
Therefore,
$p'_F - p'_M \sim N\left(0, (0.077)(0.923)\left(\frac{1}{2169} + \frac{1}{2231}\right)\right)$
$p'_F - p'_M$ follows an approximate normal distribution.

Calculate the *p*-value using the normal distribution:
p-value = 0.1045
Estimated proportion for females: 0.0719
Estimated proportion for males: 0.082

Graph:



Decision: Since $\alpha < p$-value, Do not reject *H0*

Conclusion: At the 1% level of significance, from the sample data, there is not sufficient evidence to conclude that the proportion of girls sending “sexts” is less than the proportion of boys sending “sexts.”

Press STAT. Arrow over to TESTS and press 6:2-PropZTest. Arrow down and enter 156 for x1, 2169 for n1, 183 for x2, and 2231 for n2. Arrow down to p1: and arrow to less than p2. Press `ENTER`. Arrow down to Calculate and press ENTER. The *p*-value is $P = 0.1045$ and the test statistic is $z =$ -1.256.

Researchers conducted a study of smartphone use among adults. A cell phone company claimed that iPhone smartphones are more popular with whites (non-Hispanic) than with African Americans. The results of the survey indicate that of the 232 African American cell phone owners randomly sampled, 5% have an iPhone. Of the 1,343 white cell phone owners randomly sampled, 10% own an iPhone. Test at the 5% level of significance. Is the proportion of white iPhone owners greater than the proportion of African American iPhone owners?

This is a test of two population proportions. Let W and A be the subscripts for the whites and African Americans. Then p_W and p_A are the desired population proportions.

Random variable:
$p'_W – p'_A =$ difference in the proportions of Android and iPhone users.

H_0: $p_W = p_A$ H_0: $p_W – p_A = 0$

H_a: $p_W > p_A$ H_a: $p_W – p_A > 0$

The words "more popular" indicate that the test is right-tailed.

Distribution for the test: The distribution is approximately normal:

$p_c = x_W + x_A n_W + n_A = 134+12$
$1343+232 = 0.0927$

$1 - p_c = 0.9073$

Therefore,

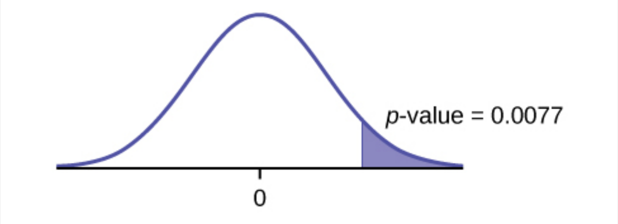
p ′ W – p ′ A ~N(0, (0.0927)(0.9073)(1 1343 + 1 232))

p ′ W – p ′ A follows an approximate normal distribution.

Calculate the *p*-value using the normal distribution:

p-value = 0.0077
Estimated proportion for group A: 0.10
Estimated proportion for group B: 0.05

Graph:



Decision: Since $\alpha > p$-value, reject the *H*0.

Conclusion: At the 5% level of significance, from the sample data, there is sufficient evidence to conclude that a larger proportion of white cell phone owners use iPhones than African Americans.

TI-83+ and TI-84: Press STAT. Arrow over to TESTS and press 6:2-PropZTest. Arrow down and enter 135 for x1, 1343 for n1, 12 for x2, and 232 for n2. Arrow down to p1: and arrow to greater than p2. Press ENTER. Arrow down to Calculate and press ENTER. The P-value is $P = 0.0092$ and the test statistic is $Z = 2.33$.

Try It

A concerned group of citizens wanted to know if the proportion of forcible rapes in Texas was different in 2011 than in 2010. Their research showed that of the 113,231 violent crimes in Texas in 2010, 7,622 of them were forcible rapes. In 2011, 7,439 of the 104,873 violent crimes were in the forcible rape category. Test at a 5% significance level. Answer the following questions:

a. Is this a test of two means or two proportions?

b. Which distribution do you use to perform the test?

c. What is the random variable?

d. What are the null and alternative hypothesis? Write the null and alternative hypothesis in symbols.

e. Is this test right-, left-, or two-tailed?

f. What is the *p*-value?

g. Do you reject or not reject the null hypothesis?

h. At the __ level of significance, from the sample data, there ____ (is/is not) sufficient evidence to conclude that ________.

References

Data from *Educational Resources*, December catalog.

Data from Hilton Hotels. Available online at http://

www.hilton.com (accessed June 17, 2013).

Data from Hyatt Hotels. Available online at http://hyatt.com (accessed June 17, 2013).

Data from Statistics, United States Department of Health and Human Services.

Data from Whitney Exhibit on loan to San Jose Museum of Art.

Data from the American Cancer Society. Available online at http://www.cancer.org/index (accessed June 17, 2013).

Data from the Chancellor’s Office, California Community Colleges, November 1994.

“State of the States.” Gallup, 2013. Available online at http://www.gallup.com/poll/125066/State-States.aspx?ref=interactive (accessed June 17, 2013).

“West Nile Virus.” Centers for Disease Control and Prevention. Available online at http://www.cdc.gov/ncidod/dvbid/westnile/index.htm (accessed June 17, 2013).

Chapter Review

Test of two population proportions from independent samples.

- Random variable: p ˆ A – p ˆ B = difference between the two estimated proportions
- Distribution: normal distribution

Formula Review

Pooled Proportion: p_c = x F + x M n F + n M

Distribution for the differences:
p ′ A − p ′ B ~N[0, p c (1− p c)(1 n A + 1 n B)]

where the null hypothesis is H_0: $p_A = p_B$ or H_0: $p_A – p_B = 0$.

Test Statistic (*z*-score): z= (p ′ A − p ′ B) p c (1− p c)(1 n A + 1 n B)

where the null hypothesis is H_0: $p_A = p_B$ or H_0: $p_A − p_B = 0$.

where

p'_A and p'_B are the sample proportions, p_A and p_B are the population proportions,

P_c is the pooled proportion, and $\boldsymbol{n_A}$ and $\boldsymbol{n_B}$ are the sample sizes.

Use the following information for the next five exercises. Two types of phone operating system are being tested to determine if there is a difference in the proportions of system failures (crashes). Fifteen out of a random sample of 150 phones with OS_1 had system failures within the first eight hours of operation. Nine out of another random sample of 150 phones with OS_2 had system failures within the first eight hours of operation. OS_2 is believed to be more stable (have fewer crashes) than OS_1.

Is this a test of means or proportions?

What is the random variable?

$P'_{OS1} – P'_{OS2}$ = difference in the proportions of phones that had system failures within the first eight hours of operation with OS_1 and OS_2.

State the null and alternative hypotheses.

What is the *p*-value?

0.1018

What can you conclude about the two operating systems?

Use the following information to answer the next twelve exercises. In the recent Census, three percent of the U.S. population reported being of two or more races. However, the percent varies tremendously from state to state. Suppose that two random surveys are conducted. In the first random survey, out of 1,000 North Dakotans, only nine people reported being of two or more races. In the second random survey, out of 500 Nevadans, 17 people reported being of two or more races. Conduct a hypothesis test to determine if the population percents are the same for the two states or if the percent for Nevada is statistically higher than for North Dakota.

Is this a test of means or proportions?

proportions

State the null and alternative hypotheses.

1. H_0: ______
2. H_a: ______

Is this a right-tailed, left-tailed, or two-tailed test? How do you know?

right-tailed

What is the random variable of interest for this test?

In words, define the random variable for this test.

The random variable is the difference in proportions (percents) of the populations that are of two or more races in Nevada and North Dakota.

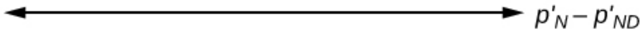
Which distribution (normal or Student's *t*) would you use for this hypothesis test?

Explain why you chose the distribution you did for the Exercise 10.56.

Our sample sizes are much greater than five each, so we use the normal for two proportions distribution for this hypothesis test.

Calculate the test statistic.

Sketch a graph of the situation. Mark the hypothesized difference and the sample difference. Shade the area corresponding to the *p*-value.



Check student’s solution.

Find the *p*-value.

At a pre-conceived $\alpha = 0.05$, what is your:

1. Decision:
2. Reason for the decision:
3. Conclusion (write out in a complete sentence):

---

1. Reject the null hypothesis.
2. *p*-value $<$ alpha
3. At the 5% significance level, there is sufficient evidence to conclude that the proportion (percent) of the population that is of two or more races in Nevada is statistically higher than that in North

Dakota.

Does it appear that the proportion of Nevadans who are two or more races is higher than the proportion of North Dakotans? Why or why not?

Homework

DIRECTIONS: For each of the word problems, use a solution sheet to do the hypothesis test. The solution sheet is found in [link]. Please feel free to make copies of the solution sheets. For the online version of the book, it is suggested that you copy the .doc or the .pdf files.

> Note
>
> If you are using a Student's *t*-distribution for one of the following homework problems, including for paired data, you may assume that the underlying population is normally distributed. (In general, you must first prove that assumption, however.)

A recent drug survey showed an increase in the use of drugs and alcohol among local high school seniors as compared to the national percent. Suppose that a survey of 100 local seniors and 100 national seniors is conducted to see if the proportion of drug and alcohol use is higher locally than nationally. Locally, 65 seniors reported using drugs or alcohol within the past month, while 60 national seniors reported using them.

We are interested in whether the proportions of female suicide victims for ages 15 to 24 are the same for the whites and the blacks races in the United States. We randomly pick one year, 1992, to compare the races. The number of suicides estimated in the United States in 1992 for white females is 4,930. Five hundred eighty were aged 15 to 24. The estimate for black females is 330. Forty were aged 15 to 24. We will let female suicide victims be our population.

---

1. H_0: $P_W = P_B$
2. H_a: $P_W \neq P_B$
3. The random variable is the difference in the proportions of white and black suicide victims, aged 15 to 24.
4. normal for two proportions
5. test statistic: –0.1944

6. *p*-value: 0.8458
7. Check student’s solution.

8.
    1. Alpha: 0.05
    2. Decision: Reject the null hypothesis.
    3. Reason for decision: *p*-value > alpha
    4. Conclusion: At the 5% significance level, there is insufficient evidence to conclude that the proportions of white and black female suicide victims, aged 15 to 24, are different.

Elizabeth Mjelde, an art history professor, was interested in whether the value from the Golden Ratio formula, (larger + smaller dimension larger dimension) was the same in the Whitney Exhibit for works from 1900 to 1919 as for works from 1920 to 1942. Thirty-seven early works were sampled, averaging 1.74 with a standard deviation of 0.11. Sixty-five of the later works were sampled, averaging 1.746 with a standard deviation of 0.1064. Do you think that there is a significant difference in the Golden Ratio calculation?

A recent year was randomly picked from 1985 to the present. In that year, there were 2,051 Hispanic students at Cabrillo College out of a total of 12,328 students. At Lake Tahoe College, there were 321 Hispanic students out of a total of 2,441 students. In general, do you think that the percent of Hispanic students at the two colleges is basically the same or different?

Subscripts: 1 = Cabrillo College, 2 = Lake Tahoe College

1. H_0: $p_1 = p_2$
2. H_a: $p_1 \neq p_2$
3. The random variable is the difference between the proportions of Hispanic students at Cabrillo College and Lake Tahoe College.
4. normal for two proportions
5. test statistic: 4.29
6. *p*-value: 0.00002
7. Check student's solution.
8.
    1. Alpha: 0.05
    2. Decision: Reject the null hypothesis.
    3. Reason for decision: *p*-value < alpha
    4. Conclusion: There is sufficient evidence to conclude that the proportions of Hispanic students at Cabrillo College and Lake Tahoe College are different.

Use the following information to answer the next three

exercises. Neuroinvasive West Nile virus is a severe disease that affects a person's nervous system . It is spread by the Culex species of mosquito. In the United States in 2010 there were 629 reported cases of neuroinvasive West Nile virus out of a total of 1,021 reported cases and there were 486 neuroinvasive reported cases out of a total of 712 cases reported in 2011. Is the 2011 proportion of neuroinvasive West Nile virus cases more than the 2010 proportion of neuroinvasive West Nile virus cases? Using a 1% level of significance, conduct an appropriate hypothesis test.

- "2011" subscript: 2011 group.
- "2010" subscript: 2010 group

This is:

1. a test of two proportions
2. a test of two independent means
3. a test of a single mean
4. a test of matched pairs.

An appropriate null hypothesis is:

1. $p_{2011} \leq p_{2010}$
2. $p_{2011} \geq p_{2010}$
3. $\mu_{2011} \leq \mu_{2010}$

4. $p_{2011} > p_{2010}$

---

a

The *p*-value is 0.0022. At a 1% level of significance, the appropriate conclusion is

1. There is sufficient evidence to conclude that the proportion of people in the United States in 2011 who contracted neuroinvasive West Nile disease is less than the proportion of people in the United States in 2010 who contracted neuroinvasive West Nile disease.
2. There is insufficient evidence to conclude that the proportion of people in the United States in 2011 who contracted neuroinvasive West Nile disease is more than the proportion of people in the United States in 2010 who contracted neuroinvasive West Nile disease.
3. There is insufficient evidence to conclude that the proportion of people in the United States in 2011 who contracted neuroinvasive West Nile disease is less than the proportion of people in the United States in 2010 who contracted neuroinvasive West Nile disease.
4. There is sufficient evidence to conclude that the proportion of people in the United

States in 2011 who contracted neuroinvasive West Nile disease is more than the proportion of people in the United States in 2010 who contracted neuroinvasive West Nile disease.

Researchers conducted a study to find out if there is a difference in the use of eReaders by different age groups. Randomly selected participants were divided into two age groups. In the 16- to 29-year-old group, 7% of the 628 surveyed use eReaders, while 11% of the 2,309 participants 30 years old and older use eReaders.

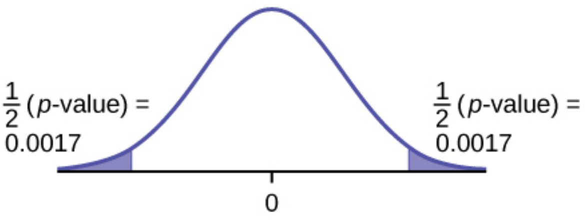
Test: two independent sample proportions.

Random variable: $p'_1 - p'_2$

Distribution:
H_0: $p_1 = p_2$
H_a: $p_1 \neq p_2$

The proportion of eReader users is different for the 16- to 29-year-old users from that of the 30 and older users.

Graph: two-tailed



p-value : 0.0033

Decision: Reject the null hypothesis.

Conclusion: At the 5% level of significance, from the sample data, there is sufficient evidence to conclude that the proportion of eReader users 16 to 29 years old is different from the proportion of eReader users 30 and older.

Adults aged 18 years old and older were randomly selected for a survey on obesity. Adults are considered obese if their body mass index (BMI) is at least 30. The researchers wanted to determine if the proportion of women who are obese in the south is less than the proportion of southern men who are obese. The results are shown in [link]. Test at the 1% level of significance.

	Number who are obese	Sample size
Men	42,769	155,525
Women	67,169	248,775

Two computer users were discussing tablet computers. A higher proportion of people ages 16 to 29 use tablets than the proportion of people age 30 and older. [link] details the number of tablet owners for each age group. Test at the 1% level of significance.

	16–29 year olds	30 years old and older
Own a Tablet	69	231
Sample Size	628	2,309

Test: two independent sample proportions

Random variable: $p'_1 - p'_2$

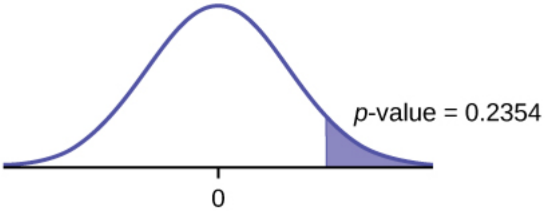
Distribution:

H_0: $p_1 = p_2$

H_a: $p_1 > p_2$

A higher proportion of tablet owners are aged 16 to 29 years old than are 30 years old and older.

Graph: right-tailed



p-value: 0.2354

Decision: Do not reject the H_0.

Conclusion: At the 1% level of significance, from the sample data, there is not sufficient evidence to conclude that a higher proportion of tablet owners are aged 16 to 29 years old than are 30 years old and older.

A group of friends debated whether more men use smartphones than women. They consulted a research study of smartphone use among adults. The results of the survey indicate that of the 973 men randomly sampled, 379 use smartphones. For women, 404 of the 1,304 who were randomly sampled use smartphones. Test at the 5% level of significance.

While her husband spent 2½ hours picking out new speakers, a statistician decided to determine whether the percent of men who enjoy shopping for electronic equipment is higher than the percent of women who enjoy shopping for electronic equipment. The population was Saturday afternoon shoppers. Out of 67 men, 24 said they enjoyed the activity. Eight of the 24 women surveyed claimed to enjoy the activity. Interpret the results of the survey.

Subscripts: 1: men; 2: women

1. H_0: $p_1 \leq p_2$
2. H_a: $p_1 > p_2$
3. $P'1 - P'2$ is the difference between the proportions of men and women who enjoy shopping for electronic equipment.
4. normal for two proportions
5. test statistic: 0.22
6. *p*-value: 0.4133
7. Check student's solution.
8.
    1. Alpha: 0.05
    2. Decision: Do not reject the null hypothesis.

3. Reason for Decision: p-value $>$ alpha
4. Conclusion: At the 5% significance level, there is insufficient evidence to conclude that the proportion of men who enjoy shopping for electronic equipment is more than the proportion of women.

We are interested in whether children's educational computer software costs less, on average, than children's entertainment software. Thirty-six educational software titles were randomly picked from a catalog. The mean cost was $31.14 with a standard deviation of $4.69. Thirty-five entertainment software titles were randomly picked from the same catalog. The mean cost was $33.86 with a standard deviation of $10.87. Decide whether children's educational software costs less, on average, than children's entertainment software.

Joan Nguyen recently claimed that the proportion of college-age males with at least one pierced ear is as high as the proportion of college-age females. She conducted a survey in her classes. Out of 107 males, 20 had at least one pierced ear. Out of 92 females, 47 had at least one pierced ear. Do you believe that the proportion of males has reached the proportion of females?

---

1. H_0: $p_1 = p_2$
2. H_a: $p_1 \neq p_2$
3. $P'1 - P'2$ is the difference between the proportions of men and women that have at least one pierced ear.
4. normal for two proportions
5. test statistic: –4.82
6. *p*-value: zero
7. Check student's solution.
8.
    1. Alpha: 0.05
    2. Decision: Reject the null hypothesis.
    3. Reason for Decision: *p*-value $<$ alpha
    4. Conclusion: At the 5% significance level, there is sufficient evidence to conclude that the proportions of males and females with at least one pierced ear is different.

Use the data sets found in [link] to answer this exercise. Is the proportion of race laps Terri completes slower than 130 seconds less than the proportion of practice laps she completes slower than 135 seconds?

"To Breakfast or Not to Breakfast?" by Richard Ayore

In the American society, birthdays are one of those days that everyone looks forward to. People of different ages and peer groups gather to mark the 18th, 20th, …, birthdays. During this time, one looks back to see what he or she has achieved for the past year and also focuses ahead for more to come.

If, by any chance, I am invited to one of these parties, my experience is always different. Instead of dancing around with my friends while the music is booming, I get carried away by memories of my family back home in Kenya. I remember the good times I had with my brothers and sister while we did our daily routine.

Every morning, I remember we went to the shamba (garden) to weed our crops. I remember one day arguing with my brother as to why he always remained behind just to join us an hour later. In his defense, he said that he preferred waiting for breakfast before he came to weed. He said, “This is why I always work more hours than you guys!”

And so, to prove him wrong or right, we decided to give it a try. One day we went to work as usual without breakfast, and recorded the time we could work before getting tired and stopping. On the next day, we all ate breakfast before going to work. We recorded how long we worked again before getting tired and stopping. Of interest was our mean increase in work time. Though not sure, my brother insisted that it was more than two hours. Using the data in [link], solve our problem.

Work hours with breakfast	Work hours without breakfast
8	6
7	5
9	5
5	4
9	7
8	7
10	7
7	5
6	6
9	5

1. H_0: $\mu_d = 0$
2. H_a: $\mu_d > 0$

3. The random variable X_d is the mean difference in work times on days when eating breakfast and on days when not eating breakfast.
4. t_9
5. test statistic: 4.8963
6. *p*-value: 0.0004
7. Check student's solution.
8.
    1. Alpha: 0.05
    2. Decision: Reject the null hypothesis.
    3. Reason for Decision: *p*-value < alpha
    4. Conclusion: At the 5% level of significance, there is sufficient evidence to conclude that the mean difference in work times on days when eating breakfast and on days when not eating breakfast has increased.

Glossary

Pooled Proportion

estimate of the common value of p_1 and p_2.